PROGRAMME AND ABSTRACTS

11th International Conference on

Computational and Financial Econometrics (CFE 2017)

http://www.cfenetwork.org/CFE2017

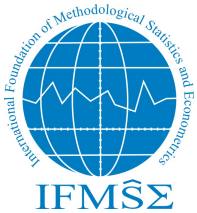
and

10th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2017)

http://www.cmstatistics.org/CMStatistics2017

Senate House & Birkbeck University of London, UK
16 – 18 December 2017



ISBN 978-9963-2227-4-2

Technical Editors: Gil Gonzalez-Rodriguez and Marc Hofmann.

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes, Marc Levene, Bernard Rachet, Herman Van Dijk.

CFE 2017 Co-chairs:
Veronika Czellar, Hashem Pesaran, Mike Pitt and Stefan Sperlich.

CFE 2017 Programme Committee:

Knut Are Aastveit, Alessandra Amendola, Josu Arteche, Monica Billio, Roberto Casarin, Gianluca Cubadda, Manfred Deistler, Jean-Marie Dufour, Ekkehard Ernst, Jean-David Fermanian, Catherine Forbes, Philip Hans Franses, Marc Hallin, Alain Hecq, David Hendry, Benjamin Holcblat, Jan Jacobs, Degui Li, Alessandra Luati, Richard Luger, J Isaac Miller, Claudio Morana, Bent Nielsen, Juan-Pablo Ortega, Michael Owyang, Gareth Peters, Cesare Robotti, Paulo Rodrigues, Jeroen Rombouts, Olivier Scaillet, Peter Siklos, Robert Taylor, Carlos Velasco, Martin Wagner and Jean-Michel Zakoian.

CMStatistics 2017 Co-chairs:
Bertrand Clarke, Steven Gilmour, Michele Guindani and Leopold Simar.

CMStatistics 2017 Programme Committee:

Peter Buehlman, Guang Cheng, Jeng-Min Chiou, Stephane Chretien, Harry Crane, Christophe Croux, Michael Daniels, Xavier De Luna, Herold Dehling, Fabrizio Durante, John Einmahl, Piotr Fryzlewicz, Yulia Gel, Armelle Guillou, Christopher Hans, Siegfried Hormann, Stephan Huckemann, Marie Huskova, Timothy Johnson, Ivan Kojadinovic, Tatyana Krivobokova, Clifford Lam, Carlos Lamarche, Christophe Ley, Marloes Maathuis, Paul McNicholas, Thomas Valentin Mikosch, Domingo Morales, Davy Paindaveine, Marco Riani, Juan Romo, Holger Rootzen, Jason Roy, Armin Schwartzman, Ana-Maria Staicu, Francisco Stingo, Stefan Van Aelst, Ingrid Van Keilegom, Tim Verdonck, Mattias Villani, Andrew T. Walden, Jane-Ling Wang, Nanny Wermuth and David van Dyk.

Local Organizer:
Birkbeck University of London.
CFEnetwork and CMStatistics.

Dear Friends and Colleagues,

We welcome you warmly to London, for the Eleventh International Conference on *Computational and Financial Econometrics* (CFE 2017) and the Tenth International Conference of the ERCIM Working Group on *Computational and Methodological Statistics* (CMStatistics 2017). As many of you know, this annual conference has become a leading joint international meeting 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 2017 programme consists of 375 sessions, 5 plenary talks and over 1550 presentations. There are about 1700 participants. Once more, this is the biggest meeting of the conference series in terms of number of participants and presentations. The growth of the conference in terms of size and quality makes it undoubtedly one of the most important international scientific events in the field.

The co-chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. The international organizing committee hopes that the conference venue will provide the appropriate environment to enhance your contacts and to establish new ones. The conference is a collective effort by many individuals and organizations. The Scientific Programme Committee, the Session Organizers, the local hosting universities and many volunteers have contributed substantially to the organization of the conference. We acknowledge their work and the support of our hosts and sponsors, and particularly Birkbeck University of London, UK.

The Elsevier journal, Econometrics and Statistics (EcoSta) has being inaugurated at the previous conference. 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.

Looking forward, the CFE-CMStatistics 2018 will be held at the University of Pisa, Italy, from Friday the 14th to Sunday the 16th of December 2018. Tutorials will take place on Thursday the 13th of December 2018. You are invited and encouraged to actively participate in these events.

We wish you a productive, stimulating conference and a memorable stay in London.

Ana Colubi, Erricos J. Kontoghiorghes and Herman K. Van Dijk: coordinators of CMStatistics & CFEnetwork.

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 1750 members and the following specialized teams

BM: Bayesian Methodology
CODA: Complex data structures and Object Data Analysis
CPEP: Component-based methods for Predictive and Exploratory Path modeling
DMC: Dependence Models and Copulas
DOE: Design Of Experiments
EF: Econometrics and Finance
GCS: General Computational Statistics WG CMStatistics
GMS: General Methodological Statistics WG CMStatistics
GOF: Goodness-of-Fit and Change-Point Problems
HDS: High-Dimensional Statistics
ISDA: Imprecision in Statistical Data Analysis
LVSEM: Latent Variable and Structural Equation Models
MCS: Matrix Computations and Statistics
MM: Mixture Models
MSW: Multi-Set and multi-Way models
NPS: Non-Parametric Statistics
OHEM: Optimization Heuristics in Estimation and Modelling
RACDS: Robust Analysis of Complex Data Sets
SAE: Small Area Estimation
SAET: Statistical Analysis of Event Times
SAS: Statistical Algorithms and Software
SEA: Statistics of Extremes and Applications
SFD: Statistics for Functional Data
SL: Statistical Learning
SSEF: Statistical Signal Extraction and Filtering
TSMC: Times Series Modelling and Computation

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.

Specialized teams

Currently the CFEnetwork has over 1000 members and the following specialized teams

AE: Applied Econometrics
BE: Bayesian Econometrics
BM: Bootstrap Methods
CE: Computational Econometrics
ET: Econometric Theory
FA: Financial Applications
FE: Financial Econometrics
TSE: Time Series Econometrics

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

2017-12-15

Welcome Reception
19:00 - 21:00

2017-12-16

Opening, 08:25
A - Keynote
CMStatistics2017
08:40 - 09:30

B
CMStatistics2017
09:40 - 10:55

Opening, 09:50
C - Keynote
CFE2017
10:05 - 10:55

Coffee Break
10:55 - 11:25

D
CFE2017 - CMStatistics2017
11:25 - 13:05

Lunch Break
13:05 - 14:35

E
CFE2017 - CMStatistics2017
14:35 - 16:15

Coffee Break
16:15 - 16:45

F
CFE2017 - CMStatistics2017
16:45 - 18:50

Christmas Conference Dinner
20:30 - 23:30

2017-12-17

G
CFE2017 - CMStatistics2017
08:40 - 10:20

Coffee Break
10:20 - 10:50

H
CFE2017 - CMStatistics2017
10:50 - 12:55

Lunch Break
12:55 - 14:25

I
CFE2017 - CMStatistics2017
14:25 - 16:05

Coffee Break
16:05 - 16:35

J
CFE2017 - CMStatistics2017
16:35 - 18:15

K - Keynote
CFE2017
18:25 - 19:15

2017-12-18

L
CFE2017 - CMStatistics2017
08:40 - 10:20

Coffee Break
10:20 - 10:50

M
CFE2017 - CMStatistics2017
10:50 - 12:55

Lunch Break
12:55 - 14:25

N - Keynote
CFE2017 - CMStatistics2017
14:25 - 15:15

O
CFE2017 - CMStatistics2017
15:25 - 16:40

Coffee Break
16:40 - 17:10

P - Keynote
CFE2017 - CMStatistics2017
17:10 - 18:00

Closing, 19:00 - 19:15

MAMMA MIA! musical
19:45 - 22:00

TUTORIALS, MEETINGS AND SOCIAL EVENTS

WINTER SCHOOL AND TUTORIALS

The COST Action CRoNoS Winter Course on Dependence models takes place on Wednesday 13th to Friday 15th December 2017 at the CLO B01 room of the Clore Management Building of Birkbeck University of London. The courses on Friday are also designated as tutorials of the conference. The first tutorial is given by Prof. Marius Hofert and Prof. Ivan Kojadinovic (Dealing with non-stationarity, serial dependence and ties in copula inference) at 9:00-13:30. The second tutorial is given by Prof. Fabrizio Durante (Tail Dependence with Copulas) at 15:00 - 19:30.

SPECIAL MEETINGS by invitation to group members

- The *CSDA Editorial Board* meeting will take place on Friday 15th of December 2017, 18:00 - 19:00, at room MAL B18 at Birkbeck. The CSDA dinner will take place on Sunday 17th of December 2017 at 20:30.
- The *Econometrics and Statistics (EcoSta) Editorial Board* meeting will take place on Sunday 17th of December 2017 from 19:15 - 20:00, at room Bloomsbury at Senate House. The EcoSta dinner will take place on Sunday 17th of December 2017 at 20:30.
- The *COST Action CRONOS* meeting will take place on Saturday 16th December 2017, 18:55-19:30, at Senate Room of the Senate House.

SOCIAL EVENTS

- *The coffee breaks* will take place at the Crush Hall and MacMillan Hall of the Senate House, at rooms MAL B04 and MAL B29 of Birkbeck University of London and at the Foyer of Clore. You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Friday 15th of December 2017, from 19:00-21:30.* The Welcome Reception is open to all registrants and accompanying persons who have purchased a reception ticket. It will take place at the Senate House Crush Hall and the MacMillan Hall (see map at page VIII). Conference registrants must bring their conference badge and ticket and any accompanying persons should bring their reception tickets in order to attend the reception. Preregistration is required due to health and safety reasons, and limited capacity of the venue. Entrance to the reception venue will be strictly allowed only to those who have a ticket.
- *Conference Dinner, Saturday 16th of December, from 20:30 to 23:30.* The conference dinner is optional and registration is required. It will take place at the hotel Galleon Suite of the Royal National Hotel (38-51 Bedford Way, London WC1H 0DG - see map at page VIII). Conference registrants and accompanying persons should bring their conference dinner tickets in order to attend the conference dinner.
- *Closing social event - MAMMA MIA!, Monday 18th of December, at 19:45.* The MAMMA MIA! musical is optional and registration is required. It will take place at the Novello Theatre (see map at page IX). Conference registrants and accompanying persons should collect their tickets from the registration desk by noon Monday the 18th December 2017. The tickets cannot be collected from the Theatre.

GENERAL INFORMATION

Addresses of venues

- Birkbeck and Clore Management Centre, University of London, Malet Street, London WC1E 7HX.
- University of London, Senate House, Malet Street, London WC1E 7HU.

Registration

The registration will be open on Friday 15th December 2017 at the Foyer of the tutorials venue, Clore Management Centre, in the morning and from 18:00 at the MacMillan Hall of the Senate House. From Saturday 16th of December to Monday 18th of December the registration will be at the McMillan Hall of the Senate House.

Lecture rooms

The paper presentations will take place at Birkbeck, Clore and Senate House (see map in the next page). Due to health and safety regulations the maximum capacity of the rooms should be respected (see the interactive programme for the details). There will be no signs indicating the location of the lecture rooms, and therefore we advise you to visit the venue in advance. The opening, keynote and closing talks will take place at the Beveridge Hall of the Senate House. The poster sessions will take place at the Macmillan Hall and Crush Hall of the Senate House.

Presentation instructions

The lecture rooms will be equipped with a PC and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide the session chair with the files for the presentation in PDF (Acrobat) or PPT (Powerpoint) format on a USB memory stick. This must be done at least ten minutes before each session. Chairs are requested to keep the sessions on schedule. Papers should be presented in the order they are listed in the programme for the convenience of attendees who may wish to go to other rooms mid-session to hear particular papers. In the case of a presenter not attending, please use the extra time for a break or a discussion so that the remaining papers stay on schedule. The PC in the lecture rooms should be used for presentations. An IT technician will be available during the conference and should be contacted in case of problems.

Posters

The poster sessions will take place at the McMillan and Crush Halls of the Senate House. The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0.

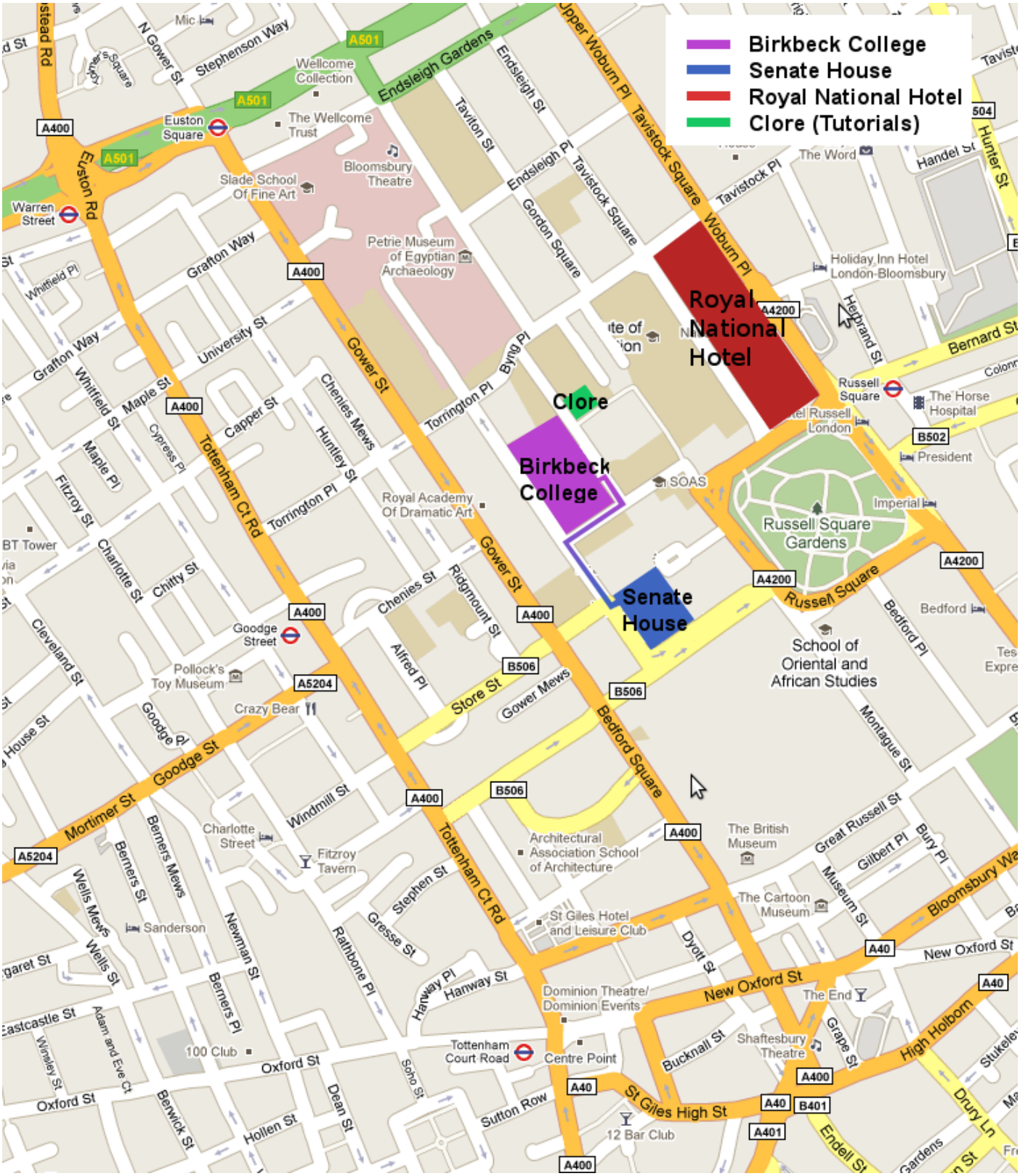
Internet Connection

Participants from any eduroam-enabled institution should use the Eduroam service in order to obtain access to Internet. For participants without Eduroam access, there will be wireless Internet connection at the Macmillan Hall. You will need to have your own laptop in order to connect to the Internet. The daily login and password will be displayed on the announcement board by the registration desk.

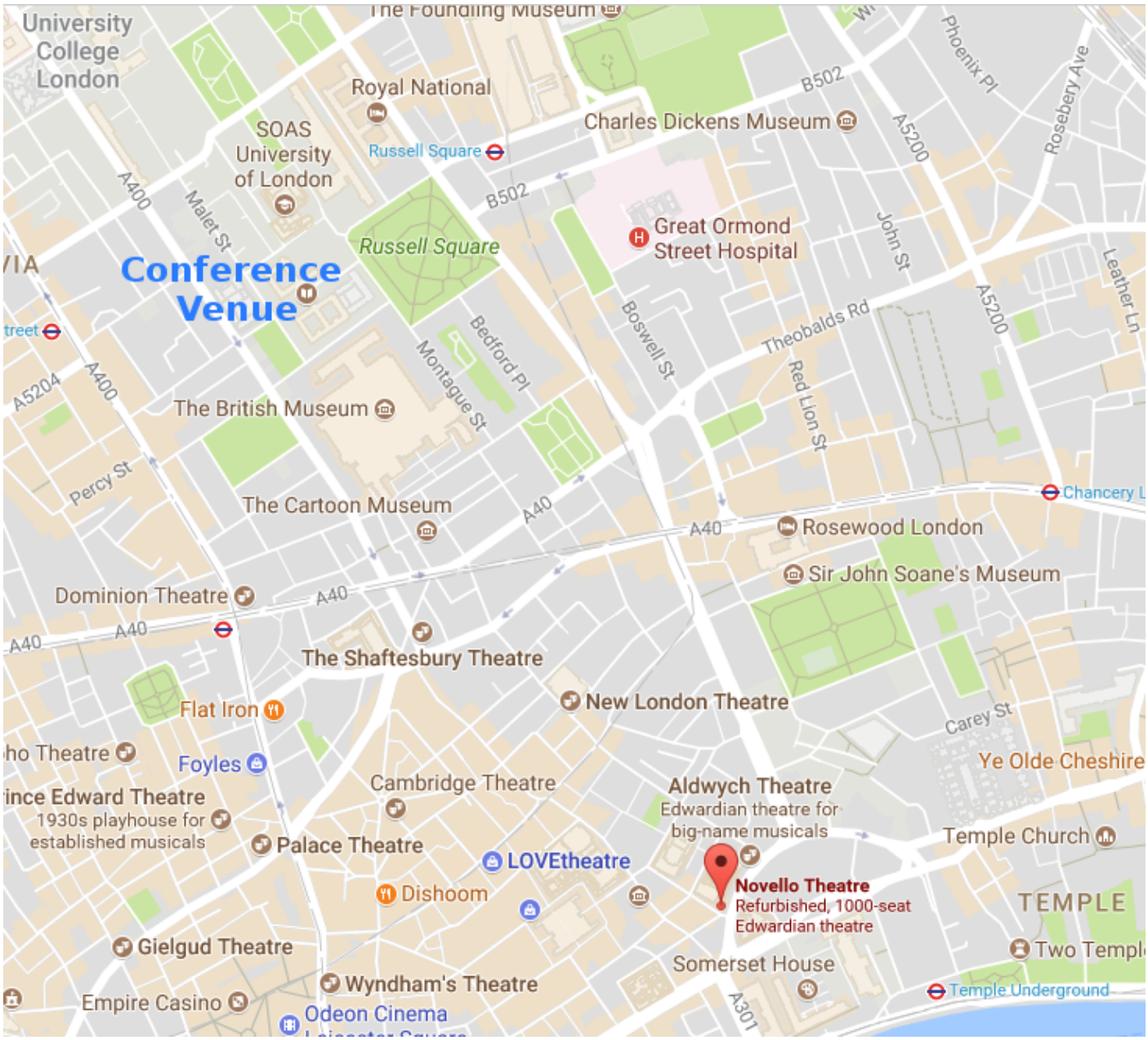
Exhibitors

Elsevier, Oxford University Press and Springer.

Map of the venue and nearby area



Extended map



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 containing novel components in econometrics and statistics are encouraged to be submitted for publication in special peer-reviewed, or regular issues of the new Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. The Econometrics and Statistics (EcoSta) is inviting submissions for the special issues with deadline for submissions the 30th April 2018:

- (Part A: Econometrics) Theoretical Econometrics.
- (Part A: Econometrics) Computational Econometrics.
- (Part B: Statistics) Copulas.
- (Part B: Statistics) Neuroimaging.

Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/ecosta (in the EES please select the appropriate special issue). For further information please consult http://www.cfenetwork.org or http://www.cmstatistics.org.

Call For Papers Computational Statistics & Data Analysis (CSDA)

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

Papers containing strong computational statistics, or substantive data-analytic elements can also be submitted to special peer-reviewed, or regular issues of the journal Computational Statistics & Data Analysis (CSDA). The CSDA is planning for 2017 the following special issues with deadline for paper submissions the 30th January 2018:

- High-dimensional and functional data analysis.
- 4th Special Issue on advances in mixture models.
- Special Issue on Biostatistics.

Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/csda (in the EES please select the appropriate special issue). Any questions may be directed via email to: csda@dcs.bbk.ac.uk.

Contents

Saturday 16.12.2017 08:40 - 09:30 Room: Beveridge Hall Chair: Michele Guindani Keynote talk 1

Piecewise deterministic Markov chain Monte Carlo for Bayesian computation

Speaker: **Arnaud Doucet, University of Oxford, United Kingdom** Alexandre Bouchard, George Deligiannidis

A novel class of non-reversible Markov chain Monte Carlo schemes relying on continuous-time piecewise deterministic Markov processes has recently emerged. In these algorithms, the state of the Markov process evolves according to a deterministic dynamics which is modified using a Markov transition kernel at random event times. These methods enjoy remarkable features including the ability to update only a subset of the state components while other components implicitly keep evolving and the ability to use an unbiased estimate of the gradient of the log-target while preserving the target as invariant distribution. However, these algorithms also suffer from important limitations. The deterministic dynamics used so far do not exploit the structure of the target. Moreover, exact simulation of the event times is feasible for an important yet restricted class of problems and, even when it is, it is application specific. We introduce novel MCMC methods addressing these shortcomings by bringing together piecewise deterministic Markov processes, Hamiltonian dynamics and slice sampling. We present novel continuous-time algorithms relying on exact Hamiltonian flows and novel non-reversible discrete-time algorithms which can exploit complex dynamics such as approximate Hamiltonian dynamics arising from symplectic integrators while preserving the attractive features of continuous-time algorithms. We demonstrate the performance of these schemes on a variety of Bayesian inference tasks.

Saturday 16.12.2017 10:05 - 10:55 Room: Beveridge Hall Chair: Gael Martin Keynote talk 2

Identification and estimation of dynamic causal effects in macroeconomics

Speaker: **Mark W Watson, Princeton University, United States** James H Stock

An exciting development in empirical macroeconometrics is the increasing use of external sources of as-if randomness to identify the dynamic causal effects of macroeconomic shocks. This approach is the time series counterpart of the highly successful strategy in microeconometrics of using external as-if randomness to provide instruments that identify causal effects. This lecture provides conditions on instruments and control variables under which external instrument methods produce valid inference on dynamic causal effects, that is, structural impulse response function; these conditions can help guide the search for valid instruments in applications. We consider two methods, a one-step instrumental variables regression and a two-step method that entails estimation of a vector autoregression. Under a restrictive instrument validity condition, the one-step method is valid even if the vector autoregression is not invertible, so comparing the two estimates provides a test of invertibility. Under a less restrictive condition, in which multiple lagged endogenous variables are needed as control variables in the one-step method, the conditions for validity of the two methods are the same.

Sunday 17.12.2017 18:25 - 19:15 Room: Beveridge Hall Chair: Jeroen Rombouts Keynote talk 3

Tests of policy ineffectiveness

Speaker: **Ron Smith, Birkbeck University of London, United Kingdom**

Tests for the effect of a policy intervention usually involve comparing post-intervention realizations of a target variable with a counterfactual: the predicted value in the absence of the policy change. Previous work will be discussed where we consider tests of this form for two cases. The first case uses a dynamic stochastic general equilibrium (DSGE) model. Here tests are unlikely to have much power unless the intervention changes the steady state, e.g. the target rate of inflation. This is because DSGE models use variables that are measured as deviations from the steady state and thus any effects of policy are transitory. The tests will also not be consistent, since as the evaluation horizon expands, the effect of the intervention declines exponentially and the power goes to zero. The second case uses reduced form or final form policy response equations, which explain the target variable by its lags, the policy variable and policy-invariant exogenous variables. Since the counterfactual is a type of forecast and since parsimonious models tend to forecast better, we may obtain more reliable estimates of the counterfactual outcomes and greater power of the test from a parsimonious policy response equation than a DSGE.

Monday 18.12.2017 14:25 - 15:15 Room: Beveridge Hall Chair: Alessandra Amendola Keynote talk 4

Model-free prediction and regression

Speaker: **Dimitris Politis, University of California, San Diego, USA**

Prediction has been traditionally approached via a model-based paradigm, i.e., (a) fit a model to the data at hand, and (b) use the fitted model in order to extrapolate/predict future data. Due to both mathematical and computational constraints, 20th century statistical practice focused mostly on parametric models. Fortunately, with the advent of widely accessible powerful computing in the late 1970s, computer-intensive methods such as the bootstrap and cross-validation freed practitioners from the limitations of parametric models, and paved the way towards the 'big data' era of the 21st century.Nonetheless, there is a further step one may take, namely going beyond even nonparametric models. The Model-Free Prediction Principle is based on the simple notion of transforming a complex dataset to one that is easier to work with, e.g., i.i.d. or Gaussian. As such, it restores the emphasis on observable quantities, i.e., current and future data, as opposed to unobservable model parameters and estimates thereof. Coupled with resampling, the Model-Free Prediction Principle allows us to go beyond point prediction in order to construct frequentist prediction intervals without resort to restrictive model assumptions. Furthermore, Model-Free Prediction ideas can be used to additionally obtain point estimates and confidence intervals for quantities of interest, leading to an alternative, transformation-based approach to statistical inference.

Monday 18.12.2017 17:10 - 18:00 Room: Beveridge Hall Chair: Berc Rustem Keynote talk 5

Linear stochastic models in discrete and continuous time

Speaker: **Stephen Pollock, University of Leicester, United Kingdom**

Statistical time series analysis commonly concerns data sampled at regular intervals from continuously varying signals. The relationships subsisting in the sampled data are usually characterised without reference to the underlying continuous signals. Nevertheless, it is sometimes desirable to attempt to reconstitute the continuous signal from the sampled data by bridging the gaps between the data points. Also, it may be required reconstruct a model of the process generating the data that represents time as a continuum rather than as a sequence of points. Techniques are described that are available for fulfilling these two objectives, and it will place particular emphasis on the correspondence between discrete and continuous models of the same process. It will be shown that the difficulties in establishing a correspondence can often be alleviated by defining a continuous-time forcing function that is bounded infrequency. In that case, there is a one-to-one correspondence between the discrete and the continuous ARMA models. An effective means of establishing a correspondence, when the forcing function is not frequency-limited, will also be demonstrated.

Saturday 16.12.2017 09:40 - 10:55 Parallel Session B – CMStatistics

EO344 Room CLO B01 BAYESIAN ANALYSIS **Chair: Miguel de Carvalho**

E0660: **Bayesian estimation of a time-varying bivariate distribution from censored data**
Presenter: **Bradley Barney**, Brigham Young University, United States
Co-authors: Garritt Page, Louise Lawson, Reese Clark, Miguel de Carvalho
Motivated by the desire to construct reference growth curves for body mass index (BMI) in preterm infants, emphasis is on estimation of a time-varying bivariate density. The two inputs to BMI calculation, weight and length, are first jointly modeled because of the lesser frequency of observing length in the available data. From this joint distribution, the BMI distribution is induced. Bayesian implementations of several methods for the bivariate estimation are considered: structured additive distributional regression with a copula, nonparametric density estimation, and quantile regression. Existing approaches are extended to adjust the estimation for censoring, reflecting a prominent feature of the collected data. The performance of these methods are assessed in a simulation study, and the results from the motivating application are presented and discussed.

E0653: **Bayesian nonparametric inference for the three-class Youden index and its associated optimal cut-points**
Presenter: **Vanda Inacio**, University of Edinburgh, United Kingdom
Co-authors: Adam Branscum
The three-class Youden index is a generalisation of the Youden index to the case where there exist three ordinal disease classes and it serves both as a measure of diagnostic accuracy and as a criterion to choose the optimal pair of cutoff values for diagnosing subjects in practice. We develop a Bayesian nonparametric approach for estimating the three-class Youden index and its associated optimal cutoff values based on Dirichlet process mixtures, which are robust priors that can handle nonstandard features of the data, such as skewness and multimodality. A simulation study is performed and an application to data concerning the Trail Making Test Part A, which has been used to assess cognitive impairment in Parkinson's disease patients, is provided.

E1015: **BASiCS: Vertical and horizontal data integration for noisy single-cell expression data**
Presenter: **Catalina Vallejos**, The Alan Turing Institute, United Kingdom
Single-cell RNA-sequencing (scRNA-seq) has transformed the field of transcriptomics, providing novel insights that were not accessible to bulk-level experiments. However, the promise of scRNA-seq comes at the cost of higher data complexity. In particular, a prominent feature of scRNA-seq experiments is strong measurement error, reflected by technical dropouts and poor correlations between technical replicates. These effects must be taken into account to reveal biological findings that are not confounded by technical variation. We will describe some statistical challenges that arise when analyzing scRNA-seq datasets. We will also introduce BASiCS (Bayesian Analysis of Single Cell Sequencing data), a Bayesian hierarchical model in which data normalization, noise quantification and downstream analyses are simultaneously performed. BASiCS exploits experimental design to disentangle biological signal from technical artifacts. This includes: (i) a vertical integration approach, where a set of technical spike-in genes is used as a gold-standard and (ii) a horizontal integration framework, where technical variation is quantified by borrowing information from multiple groups of samples. Using control experiments and case studies, we will illustrate how BASiCS goes beyond traditional differential expression analyses, identifying changes in cell-to-cell gene expression variability between pre-specified groups of cells.

EO583 Room Bloomsbury COMPLEX DATA MODELING AND COMPUTATIONAL METHODS **Chair: Mauricio Castro**

E0195: **Robust inference for ROC regression**
Presenter: **Vanda Lourenco**, Faculty of Sciences and Technology - New University of Lisbon, Portugal
Co-authors: Vanda Inacio, Miguel de Carvalho
The receiver operating characteristic (ROC) curve is the most popular tool for evaluating the diagnostic accuracy of continuous biomarkers. Often, covariate information that affects the biomarker performance is also available and several regression methods have been proposed to incorporate covariates in the ROC framework. We propose robust inference methods for ROC regression, which can be used to safeguard against the presence of outlying biomarker values. Simulation results suggest that the methods perform well in recovering the true conditional ROC curve and corresponding area under the curve, on a variety of data contamination scenarios. Methods are illustrated using data on age-specific accuracy of glucose as a biomarker of diabetes.

E0531: **Mixtures of common factor analyzers based on the restricted multivariate skew-t distribution**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan
Co-authors: Wan-Lun Wang, Mauricio Castro
Mixtures of common t factor analyzers (MCtFA) have been shown its effectiveness in robustifying mixtures of common factor analyzers (MCFA) when handling model-based clustering of the high-dimensional data with heavy tails. However, the MCtFA model may still suffer from a lack of robustness against observations whose distributions are highly asymmetric. A further robust extension of the MCFA and MCtFA models, called the mixture of common skew-t factor analyzers (MCstFA), is presented by assuming a restricted multivariate skew-t distribution for the common factors. The MCstFA model can be used to accommodate severely non-normal (skewed and leptokurtic) random phenomena while preserving its parsimony in factor-analytic representation and performing graphical visualization inlow-dimensional plots. A computationally feasible Expectation Conditional Maximization Either (ECME) algorithm is developed to carry out maximum likelihood estimation. The numbers of factors and mixture components are simultaneously determined based on common likelihood penalized criteria. The usefulness of the proposed model is illustrated with simulated and real datasets, and results signify its superiority over some existing competitors.

E0565: **Multivariate-t linear mixed models for multiple repeated measures with censored data**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan
Co-authors: Tsung-I Lin, Victor Hugo Lachos Davila
The analysis of complex longitudinal data is challenging due to several inherent features. First, more than one series of responses are repeatedly collected on each subject at irregularly occasions over a period of time. Secondly, censorship due to limits of quantification of responses arises left-, right- and/or interval- censoring effects. Thirdly, outliers or heavy-tailed noises are possibly embodied within multiple response variables. The multivariate-t linear mixed model with censored responses (MtLMMC) is presented, which allows the analysts to model such data in the presence of the above described features simultaneously. We develop an efficient expectation conditional maximization either (ECME) algorithm to carry out maximum likelihood estimation of model parameters. The implementation of the E-step relies on the mean and covariance matrix of truncated multivariate-t distributions, which can be done by using the R TTmoment package. The proposed methodology is illustrated through a simulation study and a HIV/AIDS example.

EO486 Room Chancellor's Hall MULTI-DIMENSIONAL MODELING TECHNIQUES FOR BRAIN IMAGING DATA Chair: Damla Senturk

E0897: **A simple algorithm for molecular reconstruction from cryo-electron microscopy**
Presenter: **Hua Zhou**, UCLA, United States

A simple and efficient algorithm for solving a version of the generalized Procrustes problem is described. It is motivated by the problem of reconstructing 3D molecular structure from cryo-electron microscopy.

E1525: **A Bayesian approach for multi-subject effective connectivity inference using multi-modal neuroimaging data**
Presenter: **Michele Guindani**, University of California, Irvine, United States

The use of multi-subject vector autoregressive (VAR) models for inference on effective connectivity is discussed based on resting-state functional MRI data. In particular, we discuss the use of a Bayesian variable selection approach to allow for simultaneous inference on effective connectivity at both the subject- and group-level. Furthermore, our proposal accounts for multi-modal data by integrating structural imaging information into the prior model, encouraging effective connectivity between structurally connected regions. The effectiveness of the approach is explored through simulation studies showing improved inference on effective connectivity at both the subject- and group-level, compared to currently used methods. Our motivating application is from temporal lobe epilepsy data, where we use resting-state functional MRI and structural MRI. The group-level effective connectivity we infer include both known relationships between resting-state networks, as well as relationships of potential interest for future investigation.

E1862: **Localized multivariate FPCA for analyzing dynamic coupling of the heart and brain**
Presenter: **Robert Krafty**, University of Pittsburgh, United States

Localized-variate functional principal component analysis (LVFPCA) is discussed for finding basis functions that account for most of the variability in a random multivariate process. As opposed to traditional methods, the basis functions found by LVFPCA can be both sparse among variates (i.e. is zero across an entire functional variate) and localized within a variate (i.e. nonzero only within a subinterval of a variate). LVFPCA is formulated as a rank-one based convex optimization problem with matrix L1 and block Frobenius norm based penalties, which induce localization and variate sparsity, respectively. The approach not only provides more accurate estimates of PCs when they are sparse among variates or localized, but it also provides a tool for obtaining more interpretable PCs. The method is motivated by and used to analyze the coupling of EEG and ECG spectral measures during different periods of sleep.

EO483 Room Court STATISTICAL MODELLING OF CLIMATE CHANGE AND WEATHER EVENTS Chair: Michael Wiper

E0851: **Hierarchical vine copula models for the analysis of glacier discharge**
Presenter: **Mario Gomez**, Universidad Carlos III de Madrid, Spain
Co-authors: Concepcion Ausin, Carmen Dominguez

Glaciers are considered sensors of the Global Warming. The study of their mass balance is essential to understand their future behaviour. One of the components of this mass balance is the loss of water produced by melting, also known as glacier discharge. The aim is to analyse the relationship among the glacier discharge and other meteorological variables such as temperature, humidity, solar radiation and precipitation, and to find a model that allow us to forecast future values of the glacier discharge. The multivariate distribution of these variables is divided into four cases according to the presence or not of non-zero discharge and/or non-zero precipitation, on the other hand, seasonal effects are captured by using different parameters for each season. Then, a different vine copula structure is proposed to model the multivariate and nonlinear dependence among these variables in each case/season. Moreover, we propose a hierarchical structure where we suppose that the relationships between each pair of the meteorological variables, in each case/season, is led by common hyperparameters. Finally, Bayesian inference is performed over this hierarchical structure with the help of Approximate Bayesian Computation (ABC) techniques.

E1020: **Cluster of trajectories of airborne microorganisms susceptible to colonize Antarctic soils in a climate change scenario**
Presenter: **Ana Justel**, Universidad Autonoma de Madrid, Spain
Co-authors: Lucas Fernandez Piana, Julio Rodriguez-Puerta, Marcela Svarc, Sergi Gonzalez

A new cluster algorithm is proposed for Large Multivariate Functional Data Analysis (LMFDA), where data sets combine geographic positioning variables with continuous variables of non-comparable magnitudes. Variable standardization is crucial and is done taking into account the nature of the data as functions. The method is applied to clustering trajectories and back-trajectories arriving to the Byers Peninsula, located at the western coast of the Livingston Island (South Shetland Islands, Antarctica), to establish the dispersal capability of microorganisms susceptible of colonizing newly exposed locations in a climate change scenario. Ten years of 5-day back trajectories each six hours, computed with the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT), are clustered in groups that circulate in nearby regions, regardless of the moment they pass, and with similar environmental conditions (temperature, radiation, etc.) Main air transport patterns capture seasonal differences and rare events that influence in the microorganism viability, but do not take into account the speed at which propagules move. Another set of trajectories passing over Byers Peninsula and with origin in some of the terrestrial regions near Antarctica, are considered. The two clusters combination provides a map of airborne microorganism dispersion in Antarctica to be associated to in-situ captured propagules.

E0850: **Time-varying nonstationary multivariate risk analysis using a dynamic Bayesian copula**
Presenter: **Michael Wiper**, Universidad Carlos III de Madrid, Spain
Co-authors: Concepcion Ausin, Ali Sarhadi

A time-varying risk analysis is proposed for an adaptive design framework in nonstationary conditions arising from climate change. A Bayesian, dynamic conditional copula is developed for modeling the time-varying dependence structure between mixed continuous and discrete multiattributes of multidimensional hydrometeorological phenomena. Joint Bayesian inference is carried out to fit the marginals and copula in an illustrative example using an adaptive, Gibbs Markov Chain Monte Carlo (MCMC) sampler. Posterior mean estimates and credible intervals are provided for the model parameters and the Deviance Information Criterion (DIC) is used to select the model that best captures different forms of nonstationarity over time. A fully Bayesian, time-varying joint return period is also introduced for multivariate time-dependent risk analysis in nonstationary environments. The results demonstrate that the nature and the risk of extreme-climate multidimensional processes are changed over time under the impact of climate change, and accordingly the long-term decision making strategies should be updated based on the anomalies of the nonstationary environment.

EO340 Room G11 NEW METHODS FOR HIGH-DIMENSIONAL DATA Chair: Ming-Yen Cheng

E0723: **Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference**
Presenter: **Tao Yu**, National University of Singapore, Singapore
Co-authors: Jing Qin, Pengfei Li, Hao Liu, Baojiang Chen

The augmented inverse weighting method is one of the most popular methods for estimating the mean of the response in causal inference and missing data problems. An important component of this method is the propensity score. Popular parametric models for the propensity score include the logistic, probit, and complementary log-log models. A common feature of these models is that the propensity score is a monotonic function of a linear combination of the explanatory variables. To avoid the need to choose a model, we model the propensity score via a semiparametric single-index model, in which the score is an unknown monotonic nondecreasing function of the given single index. Under this new model, the augmented inverse weighting estimator of the mean of the response is asymptotically linear, semiparametrically efficient, and more robust than existing estimators. Moreover, we have made a surprising observation. The inverse probability weighting and augmented inverse weighting estimators based on a correctly specified parametric model may have worse performance than their counterparts based on a nonparametric model. A heuristic explanation of this phenomenon is provided. A real data example is used to illustrate the proposed methods.

E1029: **A group lasso-based method with its application to biomedical spectroscopic data**
Presenter: **Ying Zhu**, Nanyang Technological University, Singapore

High dimensional spectroscopic data consist of many overlapping absorption bands sensitive to the physical and chemical states of compounds and thus show highly correlated structure due to the complex system of biomedical spectroscopic data. In certain settings, there may exist distinct groups among the variables, and it could be more informative to exploit these groups when performing classification of samples. A model based on group lasso was developed by encouraging correlated spectral features within a group to have a shared association with the response. The proposed model characteristic with grouping effect yields more interpretable results in an application to biomedical spectroscopic data. The informative spectral absorption bands selected for classification have provided great evidence regarding the bioactive constituents of the samples.

E0232: **An investigation of a generalized least squares estimator for non-linear time series models**
Presenter: **Xiaoling Dou**, Waseda University, Japan

Ochi's estimator for the autoregressive coefficient of the first-order autoregressive model (AR(1)) uses two constants for the end points of the process. Classical estimators for AR(1) , such as the least squares estimator, Burg's estimator, and Yule-Walker estimator are obtained as special cases by choice of the constants in Ochi's estimator. By writing the first-order autoregressive conditional heteroskedastic model, ARCH(1), in a form similar to that of AR(1), we extend Ochi's estimator to ARCH(1) models. This allows the introduction of analogues of the least squares estimator, Burg's estimator and Yule-Walker estimator. We compare the relations of these with Ochi's estimator for ARCH(1) models. We then provide a simulation for AR(1) models and examine the performance of Ochi's estimator. Also, we simulate Ochi's estimator for ARCH(1) with different parameter values and sample sizes.

EO746 Room Gordon ADVANCES IN CLUSTERING AND REPRESENTATION MODELS Chair: J Fernando Vera

E1852: **Cluster and structural equation multidimensional scaling for response and predictor spaces in distance-based regression**
Presenter: **Eva Boj**, University of Barcelona, Spain
Co-authors: J Fernando Vera

In cluster-based distance regression analysis, the use of cluster analysis in conjunction with MDS is an advisable procedure to reduce the number of elements to be represented using dissimilarities. Then, the projection of the vector of continuous responses in a Euclidean space of low dimensionality is given by multidimensional scaling. When two sets of variables constitute predictor and response space, clustering separately in both spaces may not be an advisable procedure for the prediction purpose. Nevertheless, two separate dissimilarity matrices measured between the observed elements in the predictor can be considered. Then, assuming both dissimilarity matrices are observations with error of an unknown symmetric dissimilarity matrix involving the overall information for the prediction and the response space, it is proposed to make clustering from the estimated values of this latent dissimilarity matrix using structural equation multidimensional scaling. From an estimated matrix of dissimilarities between the thus given clusters, distance-based regression analysis can thus be formulated between clustered elements. The performance of the proposed procedure is illustrated with the analysis of real data sets in an econometric context.

E1854: **Weighted metric scaling in logistic classification**
Presenter: **Teresa Costa**, Universitat de Barcelona, Spain
Co-authors: Eva Boj

Weighted distance-based regression was constructed by using the theory and properties of weighted metric scaling. This model is the named distance-based linear model (db-lm). Later in the literature it has been defined the distance-based generalized linear model (db-glm) which allows us to assume error distributions in the exponential family and link functions as in any generalized linear model. Db-glm is fitted using an iterative weighted least squares algorithm where db-lm substitutes ordinary linear model. In linear models, prediction error can be estimated by the squared root of the sum of the process variance and of the estimation variance. The part of the estimation variance can be approximated by applying the delta method and/or by using bootstrap. All these formulations are studied for the generalized linear model with Binomial error distribution and logit link function, the logistic regression. They are illustrated with a real data set with the aim of classifying individuals. Db-lm and db-glm can be fitted with functions dblm and dbglm of the dbstats package for R.

E1853: **Latent block distance-association model**
Presenter: **J Fernando Vera**, University of Granada, Spain

Log-lineal models with a large number of parameter to be estimated results in tables with large number of cells. Distance association models have been proposed to reduce the number of parameters and also facilitate the interpretation. For tables involving profiles, the DA model can be estimated but the given results may be difficult to interpret because the presence of a large amount of modalities and/or zeros. Although collapsing rows is an advisable procedure this procedure may still fail in the representation of associations for tables having a large number of modalities in the response variable, and in particular for sparse tables as a profile by profile sparse contingency table. A latent block distance association model is formulated that aims the simultaneous partitioning of the rows and the columns of a contingency table, while the between blocks association is represented in a low dimensional space in terms of Euclidean distances. In the LBDA model, odds are defined in terms of the block-related main effects and of the distances, while odds ratio are defined only in terms of the squared distances.

EO290 Room CLO 101 BAYESIAN DECISION AND RISK Chair: Jacinto Martin Jimenez

E0712: **A decision model for combining energy storage technologies**
Presenter: **Emilio Lopez Cano**, Universidad Rey Juan Carlos, Spain
Co-authors: Javier Martinez Moguerza, Antonio Alonso Ayuso

In recent years, energy storage came at the forefront of mainstream discussions about how to reach a global sustainable energy future. In fact, energy storage and related technologies are increasingly playing a prominent role in the global energy debate. Thus, decision models are needed at several levels in order to tackle the societal challenge of providing efficient, scalable, secure, and robust energy storage solutions. The optimization model for energy storage systems presented considers, in addition to existing technologies, emerging or forecast ones. It envisages an heterogeneous set of technologies working as silos of energy, connected to resources and final uses. Moreover, it scales from the short term to the long term, allowing to use cooperation mechanisms such as statistical transfer in order to support policy making. The model is suitable for different objectives and, hence, multiple agents' risks, that are to be managed within the own optimization model. Thus, different risk management options, including Bayesian ones, are discussed.

E1196: **Bayesian analysis of risk measures in finance**
Presenter: **Jacinto Martin Jimenez**, Universidad de Extremadura, Spain
Co-authors: Eva Sanjuan, M Isabel Parra Arevalo

Assessing the probability of rare and extreme events is an important issue in risk management of financial portfolios. The most traditional risk measures (RM) in this context are Value at Risk, Expected Shortfall and the Return Level. Usually, a flexible Generalized Pareto Distribution (GPD) is used to model the values over a certain threshold. The parameter of this distribution have an great impact on the values of the RM. Considering a Bayesian analysis we can obtain a predicted distribution for those values. This distribution gives us more information than using a point estimation of the parametres of the GPD. We illustrate the ideas and compare both approaches with several simulation examples.

E0951: **Bayesian sensitivity analysis of the parameters of a GPD using distorted band classes**
Presenter: **Jose Pablo Arias-Nicolas**, University of Extremadura, Spain
Co-authors: Alfonso Suarez-Llorens

The objective of extreme value analysis is to model and measure tail events that occur with small probability, using only extreme values above some high threshold rather than using all of the data. It is well known that, for high thresholds, the excess distribution function can be approximated by a Generalized Pareto Distribution (GPD), which is used as much more reliable than the normal distribution due to the fact that gives the accent on the extreme values. Two measures that we find most useful and reliable for describing the tail of the distribution are value-at-risk and expected shortfall. Robust Bayesian analysis, also called Bayesian sensitivity analysis, aims to quantify and interpret the uncertainty induced by the partial knowledge of one of the three elements in Bayesian analysis (prior, likelihood and loss). Studies mainly focus on computing the range of some quantities of interest when the prior distribution varies in a class. We use the band distorted class to compute the range of the parameters of a Generalized Pareto Distribution. The two risk measures, value-at-risk and expected shortfall, are constructed based on the Bayesian estimation results.

EO609 Room CLO 102 ADVANCES IN STATISTICAL MODELLING: THEORY AND APPLICATIONS Chair: Inmaculada Barranco-Chamorro

E0684: **Application of random forests and ANOVA techniques to the aggregate modeling of road accident time series in Spain**
Presenter: **Jose Mira**, Universidad Politecnica de Madrid, Spain
Co-authors: Almudena Sanjurjo de No, Camino Gonzalez, Blanca Arenas

The purpose is the application of machine learning techniques such as Random Forests to the macro (aggregate) modeling of road accident data in Spain, along the period 2004-2013. Although the number of people killed on the road has decreased dramatically in Spain in the last decade (4032 in 2003 vs 1160 in 2016), road accidents are still a major cause of death, particularly among young people, and considerable research is being carried out to further decrease the figures. Dynamic time series models have been traditionally applied to this kind of data, where the output dynamic variable is either the number of deaths or of fatal accidents, and the explanatory variables are the lags of the output and the evolution of fleets, economic variables, weather conditions and, very importantly, enforcement policies including number of agents on the road or changes in legislation. The same data have been used in a non parametric regression model such as CART-Random Forest. Also, an experimental design methodology has been developed for a more sophisticated and statistically-based input variable selection procedure.

E0877: **Modelling real phenomena with power law tail by the family of generalized power law distributions**
Presenter: **Faustino Prieto**, University of Cantabria, Spain
Co-authors: Jose Maria Sarabia

Power laws (Pareto distributions) are very common in physics and social sciences. However, they are usually valid only in the upper tail of the distribution. We explore the properties of a new family of distributions - the family of Generalized Power Law (GPL) distributions - that we could use to model, in the whole range, real phenomena with power law tail. In addition, we provide empirical evidence of the efficacy of those distributions with real datasets. To do that, we use the following methodologies: (1) maximum likelihood method for fitting the models to the datasets; (2) Bayesian information criterion for comparing them with other well-known models; (3) rank-size plot as a graphical method; (4) Kolmogorov-Smirnov test method based on bootstrap resampling for testing the goodness of fit of those models.

E0910: **Statistical tools to deal with off-diagonal elements in square asymmetric matrices obtained from spatial data**
Presenter: **Inmaculada Barranco-Chamorro**, University of Seville, Spain

Change and confusion matrices are a quite common tool used in spatial statistics. They are usually obtained as a result from classification of images and fotointerpretations. From the statistical point of view, we have a square asymmetric contingency table where the rows and columns refer to the same set of objects, and their entries are non negative integer numbers. Quite often, standard statistical methods designed for contingency tables are unsuccessful to deal with this kind of matrices because of the strong effect of the diagonal values on the statistical summaries. So, we focus on matched-pairs statistical methods to deal with the off-diagonal elements of these matrices. Marginal homogeneity tests are considered where exact and approximate results are given. Practical applications to real spatial data are included.

EO534 Room Jessel RECENT DEVELOPMENTS IN HIGH-DIMENSIONAL MODELING, INFERENCE AND COMPUTATION Chair: Yun Yang

E1254: **An iterative penalized least squares approach to sparse canonical correlation analysis**
Presenter: **Qing Mai**, Florida State University, United States

It is of increasing interest to model the relationship between two sets of measurements when both of them are high-dimensional. Canonical correlation analysis (CCA) is a classical tool that explores the dependency of two multivariate random variables and extracts canonical pairs of linear combinations that are highly correlated. Many recent studies aim to generalize the classical CCA to high-dimensional settings. However, most of the existing CCA methods either rely on strong assumptions on the covariance matrices, or do not produce nested solutions. We propose a new sparse CCA (SCCA) method that recasts high-dimensional CCA as an iterative penalized least squares problem. Thanks to the new penalized least squares formulation, our SCCA method directly penalizes and estimates the sparse CCA directions with efficient algorithms. Therefore, in contrast to some existing methods, the new SCCA does not impose any sparsity assumptions on the covariance matrices. The proposed SCCA is also very flexible in the sense that it can be easily combined with properly chosen penalty functions to perform structured variable selection or to incorporate prior information. Moreover, our proposal of SCCA produces nested solutions, which provides great convenient in practice. Theoretical results show that SCCA can consistently estimate the true canonical pairs with an overwhelming probability in ultra-high dimensions. Numerical results also demonstrate the competitive performance of SCCA.

E1046: **Bayesian model selection for semi-parametric models**
Presenter: **Weining Shen**, UC Irvine, United States

Bayesian analysis is conducted on a class of semi-nonparametric regression models with high-dimensional parametric covariates. In particular, we show (1) strong model selection consistency, where the posterior probability of the true model converges to one; and (2) joint BvM theorem, where the posterior of the selected parametric and nonparametric components jointly converges to a Gaussian vector.

E0585: **Convex tensor clustering with applications to online advertising**
Presenter: **Will Wei Sun**, University of Miami School of Business Administration, United States
Co-authors: Brian Gaines, Eric Chi, Hua Zhou

Tensors, as a multi-dimensional generalization of matrices, have received increasing attention in industry due to its success in modeling data with complex structures. One typical circumstance is in online advertising, where user click behavior on different ads from multiple publisher platforms forms a user-ad-publisher tensor. The goal is to simultaneously group users, ads, and publishers for better targeted advertising. We will discuss a convex formulation of the tensor clustering problem, which is guaranteed to obtain a unique global minimum. It generates an entire solution path of clusters in all tensor modes governed by one tuning parameter, and thus alleviates the need to specify the number of clusters a priori. The finite sample error bound of the proposed estimator reveals an interesting bless of dimensionality phenomenon in the tensor clustering. To demonstrate the potential business impact of our method, we conduct convex clustering on the user-ad-publisher tensor data obtained from a major online company. Our clustering results provide interesting insights in understanding the user click behavior.

EO234 Room Montague CAUSAL INFERENCE IN HIGH DIMENSIONAL SETTINGS Chair: Jason Roy

E0673: **High-dimensional confounding adjustment using continuous spike and slab priors**
Presenter: **Joseph Antonelli**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Giovanni Parmigiani, Francesca Dominici

In observational studies, estimation of causal effects relies on proper adjustment for confounding. If the number of the potential confounders (p) is larger than the number of observations (n), then direct control for all potential confounders is infeasible. Existing approaches for dimension reduction and penalization are for the most part aimed at predicting the outcome, and are not suited for estimation of causal effects. We propose continuous spike and slab priors on the regression coefficients β_j corresponding to the potential confounders X_j when $p \geq n$. If X_j is associated with the treatment, then we increase the prior probability that β_j is included in the slab component of the prior. This reduces the degree of shrinkage of β_j towards zero. Using theoretical considerations and a simulation study we compare our proposed approach to alternatives and show its ability to adjust for confounding across a range of data generating mechanisms. Finally, we estimate the causal effects of persistent pesticide exposure on triglyceride levels, which highlights the key important features of our approach: 1) the ability to identify the true confounders, and 2) the ability to exclude instrumental variables, therefore minimizing bias and improving efficiency of effect estimates.

E0786: **Precision medicine in high dimensional settings**
Presenter: **Ashkan Ertefaie**, University of Rochester, United States
Co-authors: Qingyuan Zhao, Dylan Small

Effect modification occurs when the effect of the treatment variable on an outcome varies according to the level of other covariates and often has important implications in decision making. When there are hundreds of covariates, it becomes necessary to use the observed data to select a simpler model for effect modification and then make valid statistical inference. A two-stage procedure is proposed to solve this problem. First, we use Robinsons transformation to decouple the nuisance parameter from the treatment effect and propose to estimate the nuisance parameters by machine learning algorithms. Next, after plugging in the estimates of the nuisance parameters, we use the Lasso to choose a sparse model for effect modification. Compared to a full model consisting of all the covariates, the selected model is much more interpretable. Compared to the univariate subgroup analyses, the selected model greatly reduces the number of false discoveries. We show that the conditional selective inference for the selected model is asymptotically valid given the classical rate assumptions in semiparametric regression. Extensive simulation studies are performed to verify the asymptotic results and an epidemiological application is used to demonstrate our method.

E0796: **Nonparametric causal effects based on incremental propensity score interventions**
Presenter: **Edward Kennedy**, Carnegie Mellon University, United States

Most work in causal inference considers deterministic interventions that set each unit's treatment to a fixed value. However, under positivity violations such interventions lead to non-identification, inefficiency, and effects with little practical relevance. Further, corresponding effects in longitudinal studies are highly sensitive to the curse of dimensionality, resulting in widespread use of unrealistic parametric models. We propose a novel solution to these problems: incremental interventions that shift propensity score values rather than set treatments to fixed values. These have several crucial advantages. First, they avoid positivity assumptions entirely. Second, they require no parametric assumptions and yet still admit a simple characterization of longitudinal effects, independent of the number of timepoints. For example, they allow longitudinal effects to be visualized with a single curve instead of lists of coefficients. After characterizing these incremental interventions and giving identifying conditions for corresponding effects, we also develop general efficiency theory, propose efficient nonparametric estimators that can attain fast convergence rates even when incorporating flexible machine learning, and propose a bootstrap-based confidence band and simultaneous test of no treatment effect. Finally we explore finite-sample performance via simulation, and apply the methods to study time-varying sociological effects of incarceration on entry into marriage.

EO262 Room Senate ROBUST STATISTICS Chair: Marco Riani

E1295: **On robust change-point detection in time series**
Presenter: **Roland Fried**, TU Dortmund University, Germany
Some robust techniques are discussed for testing whether a time series is second order stationary. This assumption is commonly applied in time series modelling. So far mainly CUSUM-type tests are applied to test this hypothesis against the alternative of a structural break. We focus on robust tests for testing whether a change of location has occurred and give some ideas how to test robustly for a change of variance or a change of autocorrelation.

E0834: **A robust proposal for functional clustering via trimming and constraints**
Presenter: **Luis Angel Garcia-Escudero**, Universidad de Valladolid, Spain
Co-authors: Agustin Mayo-Iscar, Joaquin Ortega, Diego Rivera Garcia
Many approaches can be found in the literature aimed at performing functional or curve clustering. However, the presence of (even a small fraction of) outlying curves may be extremely harmful for most of them, because they are not specifically designed to cope with contaminating curves. Taking this problem into account, a robust model-based clustering methodology is proposed. The proposed methodology relies on the "small-ball pseudo-density" approach for functional data that results in different model-based techniques. An impartial (i.e. data-driven) trimming is used to improve the associated robustness performance of these model-based approaches. Appropriate constraints on the involved scatter parameters are critical to get robustness and useful to avoid the detection of (non-interesting) spurious clusters. A computationally feasible algorithm is presented together with graphical tools aimed at making sensible choices for the corresponding tuning parameters. The procedure is illustrated in both simulated and real data sets.

E0748: **Robust joint modeling of mean and dispersion for GLMs**
Presenter: **Tim Verdonck**, KU Leuven, Belgium
Co-authors: Pieter Segaert, Stefan Van Aelst
Generalized linear models form a unified way of modeling the mean response under different distributions belonging to the exponential family. Because of their flexibility, they have become a powerful tool in statistics. Real data often show a larger or smaller variability than expected from the model and the dispersion may even change for different observations in the data. It is crucial to properly account for this dispersion. A typical problem in analysing real data is the possible presence of outliers in the data. As classical methods try to fit an optimal model for all observations they are highly susceptible to these atypical observations. Therefore we propose a robust procedure for jointly modeling the mean and dispersion under the GLM framework. Our robust double exponential estimator models both mean and dispersion behaviour based on a possibly different set of predictors. The good performance of our methodology is illustrated in a simulation study and on real data.

EO061 Room Woburn MODERN FINANCIAL NETWORKS: FROM CUSTOMERS TO INSTITUTIONS Chair: Vyacheslav Lyubchich

E1316: **Social network analysis and deep learning for customer retention in retail banking**
Presenter: **Vyacheslav Lyubchich**, University of Maryland, United States
Co-authors: Yuzhou Chen, Yulia Gel
Modern capabilities of customers analytics allow retailers to track individual clients and their peers and to study their behavioral patterns in order to enhance customer segmentation, to optimize next best offer strategies, and to predict loss of clients, or churn. While telecommunication providers have been using social peer network data to improve their customer analytics and business intelligence solutions for a number of years, there yet exists a very limited knowledge on the peer-to-peer relational effects in retail banking. We study the impact of social network variables on customer attrition in retail banking in Canada.

E0929: **Adding value to credit scoring using mobile phone data and social networks**
Presenter: **Maria Oskarsdottir**, KU Leuven, Belgium
One of the oldest applications in analytics is credit scoring where, traditionally, peoples banking history is used to assess their creditworthiness. However, as data is continuously being generated in more volume and variety than ever before, new credit assessment methods are emerging. In particular, new variables to capture borrower behavior going beyond simple repayment history have been shown to be good predictors of whether or not people will default on their loans. We show how both statistical and economic model performance of credit scoring models can be enhanced by incorporating alternative data sources. We build networks using call-detail records and extract features that represent the calling behavior of customers. In addition, we apply social network analytics techniques where we propagate influence from prior defaulters throughout the network to inspect the influence of known defaulters on others. The resulting influence features and the calling behavior features are used together with traditional bank history features when building the scorecards. According to our results, including network information in credit scoring models significantly increases their performance when measured in AUC. Furthermore, we demonstrate benefit in terms of profit by applying a profit measure, which we further expand for profit-based feature selection.

E1324: **Measuring asset holdings in the banking sector with balance sheet driven probability factorization**
Presenter: **Celso Brunetti**, Bocconi University and Federal Reserve Board, United States
The response of governments around the world to recent financial crises has been to provide regulators with access to several new financial and economic data streams with the promise that the increased data availability will lead to more effective risk monitoring. A major challenge then for regulators is to integrate these data together for new and timely insights into the potential channels for the propagation of risk that could endanger the overall financial system. We provide novel methodology that regulators can use to monitor the response of the interbank and stock markets, yielding insight into the balance sheets of banks at a higher frequency than standard disclosures allow, and complimenting other approaches that build on network science to assess systemic risk levels. The approach is based on a probability matrix factorization that is developed from a rigorous accounting framework to integrate the data sources together. Using data from before, during and after the 2008 financial crisis, we study the behavior of the banking sector to find evidence that measures derived from our approach leads multiple risk related variables published by the European Central Bank.

EO264 Room CLO 203 TESTING IN COMPLEX PROBLEMS Chair: Cristina Butucea

E1375: **Multidimensional two-component Gaussian mixtures detection**
Presenter: **Clement Marteau**, Universita Lyon 1, France
Let $(X_1,\ldots,X_n)$ be a d-dimensional i.i.d sample from a distribution with density f. The problem of detection of a two-component mixture is considered. The aim is to decide whether f is the density of a standard Gaussian random d-vector ($f=\phi_d$) against f is a two-component mixture. Optimal separation conditions on the mixture parameters and the dimension d are established, allowing to separate both hypotheses with prescribed errors. Several testing procedures are proposed and two alternative subsets are considered. The inverse problem and direct problem point of view will also be briefly discussed.

E1492: **Nonparametric goodness-of-fit test of the baseline intensity in the Cox model**
Presenter: **Uriel Lecleire**, LAMA, France
Co-authors: Cristina Butucea, Thierry Jeantheau, Emilie Gerardin

The right-censored times X_i, their failure indicators δ_i, and a vector of p covariates Z^i are observed for n independent individuals. The hazard rate for the uncensored times is $h(t|Z) = h(t)exp(g(Z))$, with h a non specified baseline intensity function. We consider the goodness-of-fit test that has a simple null hypothesis $h \equiv h_0$ for some given function h_0. The alternative is expressed in a nonparametric setup and separated from h_0 in a weighted $\mathbb{L}_2$ norm. We use a projection on a orthonormal basis of this weighted $\mathbb{L}_2$ space to estimate the distance and build a statistic for the goodness-of-fit test. As the weight heavily depends on the distribution of the censored times and of the covariates, as well as on the function g in our model, we study the influence of plugging-in estimators of these functions on the error probabilities of our test procedure. We show that our rates for testing are faster than the rates for estimation in the same model. Thus, we provide a nonparametric test procedure for the baseline intensity function in the Cox model.

E1799: **Testing a covariate dependent model for first-order intensity**
Presenter: **Maria Isabel Borrajo**, University of Santiago de Compostela, Spain
Co-authors: Wenceslao Gonzalez-Manteiga, Lola Martinez-Miranda

Modelling the first-order intensity function in one of the main aims in point process theory, and it has been approached so far from different perspectives. One appealing model describes the intensity as a function of a spatial covariate. In the recent literature, estimation theory and several applications have been developed assuming this hypothesis, but without formally checking the goodness-of-fit of the model. We address this problem and test whether the model is appropriate. We propose a test statistic based on a L^2-distance. We prove the asymptotic normality of the statistic and suggest a bootstrap procedure to calibrate the test. We present two applications with real data and a simulation study to better understand the performance of our proposals.

EO222 Room CLO 204 SPREADING OUT THE OPTIMAL DESIGN OF EXPERIMENTS Chair: Victor Casero-Alonso

E0249: **Bayesian analysis of data from experiments subject to restricted randomisation**
Presenter: **Sadiah Aljeddani**, University of Southampton, United Kingdom
Co-authors: Kalliopi Mylona, David Woods

Designs with restricted randomisation are often more cost-effective than completely randomised designs. They are popular in industry both for developing new products and processes and for quality improvement experiments. Model and variable selection for such experiments is crucial. We study the performance of Bayesian variable selection for analysing data from experiments with restricted randomisation using shrinkage priors that allow simultaneous model selection and parameter estimation. The properties of these procedures are illustrated via simulation studies and the analysis of exemplar data sets.

E1162: **Exact c-optimal designs for models with correlated observations**
Presenter: **Juan M Rodriguez-Diaz**, University of Salamanca, Spain

In the field of optimal design of experiments, different optimality criteria can be considered depending on the objectives of the practitioner. c-optimality is one of the most used, since it looks for the best design for estimating the linear combination of the model parameters given by vector c. The procedure for computing c-optimal approximate designs for independent observations is adapted here for the correlated setup and exact designs, that may contain repeated points and thus produce singular covariance matrices.

E0856: **Optimal design for detecting hormesis**
Presenter: **Victor Casero-Alonso**, University of Castilla-La Mancha, Spain
Co-authors: Weng Kee Wong, Andrey Pepelyshev

Hormesis has been widely observed and debated in a variety of context in the biological and toxicological sciences. Detecting its presence is an important problem with wide ranging implications. While such issues have been well discussed in the literature across disciplines, there is very little work on constructing an efficient experiment to detect its existence. We develop optimal designs for detecting hormesis for additional previously proposed models for human and ecological risk assessment. To facilitate experimenters use of more efficient designs in practice, we implement codes on a user-friendly web site to freely generate and compare different types of optimal designs tailored made to their problem. The online tool also enables the user to evaluate robustness properties of a selected to various assumptions before implementation.

EO575 Room SH349 MODEL-BASED CLUSTERING Chair: Vincent Vandewalle

E1138: **Total sum of squares decomposition for mixtures of regressions**
Presenter: **Salvatore Ingrassia**, University of Catania, Italy
Co-authors: Antonio Punzo

The aim is to propose a three-term decomposition of the total sum of squares, and a related coefficient of determination, for mixtures of (linear) regressions whose parameters are estimated by maximum likelihood, via the expectation-maximization algorithm, under normally distributed errors in each mixture component. In particular, three types of mixtures of regressions are considered: with fixed covariates, with concomitant variables, and with random covariates. A ternary diagram is also suggested to make easier the joint interpretation of the three terms of the proposed decomposition. Two applications to real data illustrate the usefulness of our decomposition

E1139: **Variable selection for model-based clustering: Application in human population genomics**
Presenter: **Matthieu Marbac**, ENSAI, France
Co-authors: Mohammed Sedki

Two approaches are considered for selecting variables in latent class analysis. The first approach consists in optimizing the BIC with a modified version of the EM algorithm. This approach simultaneously performs both model selection and parameter inference. The second approach consists in maximizing the MICL, which considers the clustering task, with an algorithm of alternate optimization. This approach performs model selection without requiring the maximum likelihood estimates for model comparison, then parameter inference is done for the unique selected model. Thus, both approaches avoid the computation of the maximum likelihood estimates for each model comparison. Moreover, they also avoid the use of the standard algorithms for variable selection which are often suboptimal (e.g. stepwise method) and computationally expensive. The case of data with missing values is also discussed. The interest of both proposed criteria is shown on an application in human population genomics problem. Data set describes 1300 patients by 160000 variables.

E1226: **Model-based co-clustering of high-dimensional functional data**
Presenter: **Faicel Chamroukhi**, Caen University, Lab of Mathematics LMNO, France

High dimensional data clustering is an increasingly interesting topic in the statistical analysis of heterogeneous large-scale data. We consider the problem of clustering heterogeneous high-dimensional data where the individuals are described by functional variables which exhibit a dynamical

longitudinal structure. We address the issue in the framework of model-based co-clustering and propose the functional latent block model (FLBM). The introduced FLBM model allows to simultaneously cluster a sample of multivariate functions into a finite set of blocks, each block being an association of a cluster over individuals and a cluster over functional variables. Furthermore, the homogeneous set within each block is modeled with a dedicated latent process functional regression model which allows its segmentation according to an underlying dynamical structure. The proposed model allows thus to fully exploit the structure of the data, compared to classical latent block clustering models for continuous non functional data, which ignores the functional structure of the observations. The FLBM can therefore serve for simultaneous co-clustering and segmentation of multivariate non-stationary functions. We propose a variational expectation-maximization (EM) algorithm (VEM-FLBM) to monotonically maximize a variational approximation of the observed-data log-likelihood for the unsupervised inference of the FLBM model.

EG022 Room G3 CONTRIBUTIONS IN LATENT VARIABLES Chair: Irini Moustaki

E1606: **Identifiability of binary Bayesian networks with one latent variable**
Presenter: **Hisayuki Hara**, Doshisha University, Japan

Parameter identifiability of graphical models with latent variables is a challenging problem. For Gaussian directed graphical models with one latent variable, some useful sufficient conditions for a model to be generically identifiable up to sign change are known. We discuss binary graphical models defined by directed acyclic graphs with one latent variable which is parental to all observable variables. We show that a part of previous results is applicable also to binary graphical models, and provide a sufficient condition for a model to be generically identifiable up to label swapping of a latent variable.

E1693: **A two-stage estimation procedure for nonlinear structural equation models**
Presenter: **Klaus Holst**, Maersk Transport and Logistics, Denmark
Co-authors: Esben Budtz-Jorgensen

Applications of structural equation models (SEMs) are often restricted to linear associations between variables. Maximum likelihood (ML) estimation in nonlinear models may be complex and require numerical integration. Furthermore, ML inference is sensitive to distributional assumptions. We introduce a simple two-stage estimation technique for estimation of nonlinear associations between latent variables. Here both steps are based on fitting linear SEMs: first a linear model is fitted to data on the latent predictor and terms describing the nonlinear effects are predicted. In the second step, the predictions are included in a linear model for the latent outcome variable. We show that this procedure is consistent and identifies its asymptotic distribution. We also illustrate how this framework easily allows the association between latent variables to be modelled using restricted cubic splines and we develop a modified estimator which is robust to non-normality of the latent predictor. In a simulation study, we compare the proposed method to ML-analysis and a simple two-stage least squares technique.

E1682: **Indirect pairwise fitting of latent autoregressive and moving average models**
Presenter: **Xanthi Pedeli**, Ca Foscari University, Italy
Co-authors: Guido Masarotto, Cristiano Varin

Several models for non-normal time series are built on a latent autoregressive and moving average component. Although these models have attractive properties and interpretation, the fitting is often difficult because computation of their likelihoods involves high-dimensional integration. To overcome the complexity of the exact likelihood one can rely on the pairwise likelihood of order m using only bivariate densities of observations separated at most by m units. However, a wrong choice of the order of the pairwise likelihood may have a significant impact on the fit of the underlying model. For example, even though the pairwise likelihood of order p is efficient for inference in the autoregressive model of order p, it is grossly inefficient for inference in moving average models. Based on the idea of the pairwise likelihood, an indirect fitting of latent autoregressive and moving average models is suggested. The estimation strategy consists of two ingredients. The first ingredient is parameter augmentation to split maximization of the pairwise likelihood in m separate sub-problems, thus reducing the computational cost of the fitting algorithm without affecting the statistical efficiency. The second ingredient is indirect inference to cope with the performance loss of moving average models using a truncated autoregressive representation.

EG028 Room G5 CONTRIBUTIONS IN EXTREME VALUES AND RARE EVENTS Chair: Laurent Gardes

E1744: **Understanding variability of models' time-averaged predictive skill in earthquake forecasting**
Presenter: **Alessia Caponera**, Sapienza, Italy
Co-authors: Maximilian Werner

The reproducibility of results across experiments is a fundamental pillar of science. A number of recent studies have failed to reproduce, however, the results of several landmark studies in psychology, biomedicine, economics, and geophysics. A factor contributing to irreproducibility may be the non-ergodicity of some systems: when the time-average does not equal the ensemble average, classical statistical inference procedures face challenges. We investigate the reproducibility of the predictive skill of probabilistic models that forecast the spatio-temporal evolution of earthquake sequences. Using time-dependent forecasts created during the 2010-2012 M7.2 Darfield (New Zealand) earthquake sequence, we compare the variability of models time-averaged predictive skill with commonly assumed ensemble confidence bounds. To understand the variability of predictive skill across multiple earthquake sequences, we simulate long earthquake catalogues from a widely-used earthquake clustering model (a Hawkes point process) and assess the convergence of the time-averaged predictive skill to the ensemble average. Our results suggest that many earthquake sequences are required before the temporal average converges to the ensemble average. Commonly used uncertainty estimates of the predictive skill are too optimistic, and therefore apparently significant differences between model performance may not indicate future performance.

E1591: **Identification of a structural break in the serial dependence of a time series of extremes**
Presenter: **Debbie Dupuis**, HEC Montreal, Canada
Co-authors: Luca Trapin

The analysis of extreme values in a stationary time series requires careful consideration of any dependence appearing in the extremes of the series. Thus, it is important to detect any changes in the temporal dependence of extreme values. We work within a hierarchical trawl-based framework to establish the first test for a change-point in the extremal dependence of a stationary time series. The test is flexible, easy to use, and can be extended along several lines. The asymptotic distributions of our estimators and our test statistics are established. A large simulation study verifies the good finite sample properties. We apply the test to surface level ozone data and show that there has been an increase in the serial dependence of their extreme levels in recent years.

E1676: **ABC model selection for spatial max-stable models applied to South Australian maximum temperature data**
Presenter: **Markus Hainy**, Queensland University of Technology, Australia
Co-authors: Xing Ju Lee, Christopher Drovandi, Anthony Pettitt

Max-stable processes are a common choice for modelling spatial extreme data as they arise naturally as the infinite-dimensional generalisation of multivariate extreme value theory. Statistical inference for such models is complicated by the intractability of the multivariate density function

in many cases. Among others, simulation-based approaches using approximate Bayesian computation (ABC) have been employed for estimating parameters of max-stable models. ABC algorithms rely on the evaluation of discrepancies between model simulations and the observed data rather than explicit evaluations of computationally expensive or intractable likelihood functions. The use of an ABC method to perform model selection for max-stable models is explored. The ABC summary statistics are selected in a semi-automatic way. Four max-stable models are considered: the extremal-t model with either a Whittle-Matern, Cauchy or powered exponential covariance function, and the Brown-Resnick model. The method is applied to annual maximum temperature data from 25 weather stations dispersed around South Australia.

Saturday 16.12.2017 11:25 - 13:05 Parallel Session D – CFE-CMStatistics

EI015 Room CLO B01 PREDICTION, PRESENT AND FUTURE **Chair: Dimitris Politis**

E0157: **Predictive consistency**
Presenter: **Philip Dawid**, University of Cambridge, United Kingdom
Co-authors: Ambuj Tewari
The aim is to model a stream of data arriving sequentially, but making no assumptions of independence, identical distribution, etc. Many standard concepts such as parametric consistency are not wholly adequate in this context, and need to be replaced by suitable predictive versions. We describe the concept of prequential (predictive sequential) consistency, and show how it holds more generally than parametric consistency. Extension is made to deal with misspecified models, and connections are drawn with statistical learning theory, and prediction with expert advice.

E0158: **Novel adaptive algorithms for prediction of individual sequences**
Presenter: **Alexander Rakhlin**, University of Pennsylvania, United States
The focus is on online prediction with covariates (online supervised learning) with data-dependent regret bounds in terms of empirical Rademacher averages. Such performance measures are known to be optimal in batch learning with i.i.d. data. Surprisingly, we are able to achieve these results for individual sequences. The development relies on a fundamental result about equivalence of decoupling inequalities for martingales and existence of certain special functions. Our algorithms are efficient whenever these special functions can be computed.

E0159: **A predictive approach to statistical problems with multiplicity**
Presenter: **Fumiyasu Komaki**, The University of Tokyo, Japan
Appropriate adjustments for multiplicity become essential in various statistical problems. Bayesian prediction based on parametric models with multiplicity is investigated. Priors for unknown parameters are constructed by using conditional mutual information between future observables and unknown parameters given observations. The priors depend not only on the parametric model of observed data but also on the choice of target variables to be predicted. The priors are often quite different from the Jeffreys priors or other widely used objective priors. In principle, our approach can be applied to various statistical problems with multiplicity by formulating them as prediction problems.

EO298 Room MAL B18 GRAPHICAL MARKOV MODELS I **Chair: Giovanni Marchetti**

E0911: **Undirected, indirected and regression graph models for categorical data in a common framework**
Presenter: **Alberto Roverato**, University of Bologna, Italy
Co-authors: Luca La Rocca
The problem of specifying a suitable parameterization for graphical models for categorical data is considered. We focus on three of the most relevant families of graphical models, that is, undirected, bidirected and regression graph models. In this respect, we first give some general properties concerning conditional independence and Moebius inversion. Next, we exploit these basic results to provide a unified approach to the parameterization of the three classes of models. The parameterizations are derived by applying, in the three cases, the same Moebius inversion formula to obtain a log-linear expansion of certain probabilities. In the undirected case this procedure leads to the usual corner-constrained parameterization of the class of log-linear models for contingency tables. This is the standard parameterization of undirected graph models, and we show that some well-known properties of this parameterization, such as the connection between vanishing terms and independence relationships, as well as the capability of defining context specific independences, follow directly from the constructing procedure. In this way, the former properties automatically hold true also for other parameterizations based on the same constructing procedure, and we exploit this feature to present the theory of the three classes of models in a common framework.

E0436: **Precision matrix estimation in large coloured graphical Gaussian models**
Presenter: **Helene Massam**, York University, Canada
Distributed estimation methods have recently been used to compute the maximum likelihood estimate of the precision matrix for large graphical Gaussian models. Our aim is to give a Bayesian estimate of the precision matrix for large graphical Gaussian models with, additionally, symmetry constraints imposed by an underlying graph which is coloured. We take the sample posterior mean of the precision matrix as our estimate. We study its asymptotic behaviour under the regular asymptotic regime when the number of variables p is fixed and under the double asymptotic regime when both p and n grow to infinity. We show, in particular, that when the number of parameters of the local models is uniformly bounded, the standard convergence rate of our estimate of the precision matrix to its true value, in the Frobenius norm, compares well with the rates in the current literature for the maximum likelihood estimate.

E1291: **Model selection and local geometry**
Presenter: **Robin Evans**, University of Oxford, United Kingdom
Model selection is a task of fundamental importance in statistics, and advances in high-dimensional model selection have been one of the major areas of progress over the past 20 years. Examples include covariate selection in linear regression, and models based on patterns of zeros in the inverse covariance matrix. Much of this progress has been due to penalized methods such as the lasso, and efficient methods for solving the relevant convex optimization problems. However in other classes, such as directed graphical models, correct model selection is provably hard. We give a geometric explanation for why standard convex penalized methods cannot be adapted to directed graphs, based on the local geometry of the different models at points of intersection. These results also show that it is 'statistically' hard to learn these models, and that much larger samples will typically be needed for moderate effect sizes. This has implications for other types of graphical model selection, and especially for causal models, as well as time series models. We provide some relevant heuristics that give insights into the feasibility of model selection in various classes of graphical model, including ancestral graph models, LWF chain graph models, and nested Markov models.

E1306: **Learning stable graphical models**
Presenter: **Sofia Massa**, University of Oxford, United Kingdom
Ongoing advances in model selection techniques for graphical models are trying to capture the structure of complex, high-dimensional datasets. Sparsity is usually invoked and techniques based on regularization, cross-validation, resampling and shrinkage estimation are becoming quite standard. One practical challenge in many applied contexts is how to assess the stability of different dependency structures and how to report the uncertainty associated with them. We will look at possible stability and uncertainty measures for undirected and chain graphs models.

EO457 Room MAL B20 SPATIAL AND MULTIVARIATE EXTREMES Chair: Emeric Thibaud

E0341: **Linear factor copula models and their properties**
Presenter: **Pavel Krupskiy**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Marc Genton
A special case of factor copula models with additive common factors and independent components is considered. These models are flexible and parsimonious with $O(d)$ parameters where d is the dimension. The linear structure allows one to obtain closed form expressions for some copulas and their extreme-value limits. These copulas can be used to model data with strong tail dependencies, such as extremes data. We study the dependence properties of these linear factor copula models and derive the corresponding limiting extreme value copulas with a factor structure. We show how parameter estimates can be obtained for these copulas and apply one of these copulas to analyze a financial data set.

E0827: **Equivalent representations of max-stable processes via ℓ^p norms**
Presenter: **Marco Oesting**, University of Siegen, Germany
Max-stable processes are popular models in spatial and spatio-temporal extremes. While they are typically written as pointwise maxima over an infinite number of stochastic processes, a family of novel representations based on ℓ^p norms is presented. This family includes both the construction of the Reich-Shaby model and the classical spectral representation by de Haan as special cases. As different representations may be advantageous for the purposes of statistical inference or simulation, formulae to switch between different equivalent representations are provided. Further, several properties of the represented processes are discussed.

E0620: **Determining the dependence structure of multivariate extremes**
Presenter: **Emma Simpson**, Lancaster University, United Kingdom
Co-authors: Jenny Wadsworth, Jonathan Tawn
When modelling multivariate extremes, it is important to consider the asymptotic dependence properties of the variables. This can have a complicated structure, with only certain subsets of variables taking their largest values simultaneously. We will discuss a method that, given a set of data, aims to establish this asymptotic dependence structure. It is a common method in multivariate extreme value analysis to consider variables in terms of their radial and angular components $(R, \mathbf{W})$. In this case, the angular component $\mathbf{W}$ takes values in the unit simplex and, conditioning on R being extreme, the position of mass on this simplex can reveal the asymptotic dependence structure of the variables of interest. In the bivariate case, this corresponds to deciding whether or not there is mass on the interior or the edges of the simplex, but in the d-dimensional case there are $2^d - 1$ sub-simplices that could contain mass. In reality, data will not lie exactly on the sub-simplices of the angular unit simplex, so assessing the asymptotic dependence structure is not a straightforward task. By partitioning the angular simplex, we aim to find the conditional probability that a point lies in a certain section given that it is extreme in terms of its radial component. This allows us to determine the asymptotic dependence structure of the variables, as well as the proportion of mass on each sub-simplex asymptotically.

E0764: **Some negative results on extreme multivariate quantiles defined through convex optimisation**
Presenter: **Gilles Stupfler**, The University of Nottingham, United Kingdom
Co-authors: Stephane Girard
A discussion of some general properties that a notion of extreme multivariate quantile should satisfy will be given. We will then recall the concept of geometric quantile by transposing the definition of a univariate quantile as a minimiser of a cost function based on the so-called check function to the multivariate case. We shall then argue that extreme versions of these geometric quantiles are not suitable for the extreme-value analysis of a multivariate data set. A particular reason for this is that when the underlying distribution possesses a finite covariance matrix then the magnitude of these quantiles grows at a fixed rate that is independent of the distribution. We shall also discuss an extension of this negative result to the wider class of geometric M-quantiles.

EO415 Room MAL B30 ADVANCES IN ANALYSIS OF CORRELATED SURVIVAL DATA Chair: Liming Xiang

E0855: **Hybrid quantile regression estimation for time series models with conditional heteroscedasticity**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong
It is well-known that financial time series display conditional heteroscedasticity. Among the large number of conditional heteroscedastic models, the GARCH process is the most popular and influential one. So far, feasible quantile regression methods for this task have been confined to a variant of the GARCH model, the linear GARCH model, owing to its tractable conditional quantile structure. The widely used GARCH model is considered. An easy-to-implement hybrid conditional quantile estimation procedure is developed based on a simple albeit nontrivial transformation. Asymptotic properties of the proposed estimator and statistics are derived, which facilitate corresponding inferences. To approximate the asymptotic distribution of the quantile regression estimator, we introduce a mixed bootstrapping procedure, where a time-consuming optimization is replaced by a sample averaging. Moreover, diagnostic tools based on the residual quantile autocorrelation function are constructed to check the adequacy of the fitted conditional quantiles. Simulation experiments are carried out to assess the finite-sample performance of the proposed approach. The favorable performance of the conditional quantile estimator and the usefulness of the inference tools are further illustrated by an empirical application.

E1007: **Joint regression analysis for survival data in the presence of two semi-competing risks**
Presenter: **Mengjiao Peng**, Nangyang Technological University, Singapore
Unlike the usual semi-competing risks setting that includes two types of events: the nonterminal and terminal events, in some clinical studies with complex disease process individuals may be at risk of several different types of clinical events. We consider a situation in which three events: two nonterminal events and one terminal event are of interest. This is the case of our motivating bone marrow transplant study in which individuals may experience the acute Graft-Versus-host disease (GvHD), relapse and death after an allogeneic transplant, where the GvHD is associated with the relapse free survival, both the GVHD and relapse are intermediate nonterminal events subject to dependent censoring by the informative terminal event death, but not vice versa. We propose a novel statistical approach that jointly models times to these three types of events using a pair of copulas to account for dependence structures, while each marginal distribution of the event times is formulated by a Cox proportional hazards model. We develop an estimation procedure based on pseudo likelihood and carry out simulation studies to examine the performance of the proposed method in finite samples. The practical utility of the proposed methodology is further illustrated with data from the motivating example on bone marrow transplant.

E0843: **Conditional modeling of survival data with semi-competing risks**
Presenter: **Tonghui Yu**, Nanyang Technological University, Singapore, Singapore
In recent years, various approaches have been developed for analysis of survival data in the presence of semi-competing risks arising in many clinical studies, where individuals are likely to experience two types of events: nonterminal and terminal. Most existing approaches use copula models to capture the dependence structure between both events. However, predicting overall survival (i.e., terminal event time) and accounting for the progression of the intermediate nonterminal event is often of interest in practice. It endeavors to aggregate the censored nonterminal event time with other risk factors for better modeling the overall survival. We propose a new semiparametric model for the terminal event time based

on proportional hazards regression conditioning over the nonterminal event time. We develop a two-stage estimation procedure and establish asymptotic properties of the proposed estimator. Simulation studies and a real data example are used to evaluate the performance of the proposed model.

E0625: **Missing covariate data in generalized linear mixed models with distribution-free random effects**
Presenter: **Liming Xiang**, Nanyang Technological University, Singapore
Co-authors: Li Liu
Generalized linear mixed models are considered in which the random effects are free of parametric distributions and missing at random data are present in some covariates. To overcome the problem of missing data, we propose two novel methods: a penalized conditional likelihood method relying on the auxiliary variable that is independent of random effects, and a two-step procedure consisting of a pairwise likelihood for estimating fixed effects in the first step and a penalized conditional likelihood for estimating random effects in the second step, relying on the auxiliary variable that is associated with random effects. Our methods require no distribution assumption for random effects and allow a nonparametric error structure, thus providing great flexibility in capturing a board range of behaviours of both the error term and random effects. We show that the proposed estimators enjoy desirable properties such as consistency and asymptotically normality, and assess their finite sample performance through extensive simulation studies. The proposed methods are further illustrated using a longitudinal data set on forest health monitoring.

EO617 Room MAL B33 STATISTICAL ANALYSIS IN FINITE AND INFINITE DIMENSIONAL HILBERT SPACES Chair: Karel Hron

E1128: **Testing a functional regression model through consistent bootstrap procedures**
Presenter: **Gil Gonzalez-Rodriguez**, University of Oviedo, Spain
Co-authors: Ana Colubi
Hilbert spaces are frequently used in statistics as a framework to deal with general random elements, specially with functional-valued random variables. The scarcity of common parametric distribution models in this context makes it important to develop non-parametric techniques, and among them, bootstrap has already proved to be specially valuable. A methodology to derive consistency results for some usual bootstrap methods when working in separable Hilbert spaces has been developed recently. By applying the proposed methodology, the consistency of well-known bootstrap procedures involving the sample mean as the naive bootstrap, bootstrap with arbitrary sample size, wild bootstrap, and more generally, weighted bootstrap methods, including double bootstrap and bootstrap generated by deterministic weights with the particular case of delete-h jackknife is proved. The aim is to illustrate how to employ the approach in the context of a functional regression hypothesis test.

E1032: **Kriging for Hilbert data over complex domains through random domain decomposition**
Presenter: **Alessandra Menafoglio**, Politecnico di Milano, Italy
Co-authors: Piercesare Secchi, Giorgia Gaetani
The analysis of complex data distributed over large or highly textured regions poses new challenges for spatial statistics. Indeed, methods to deal with spatial data often rely upon global models for the dependence of the field, and are thus hardly usable in the presence of textured or convoluted domains, with holes or barriers. We propose a novel methodology for the spatial prediction of Hilbert data distributed in such kinds of domains, which cope with the data and domain complexities through a divide-et-impera approach. As a key element of innovation, we propose to perform repeated Random Domain Decompositions (RDDs), each defining a set of homogeneous sub-regions where to perform locally stationary analyses. We account for the complexity of the domain by defining these partitions through appropriate non-Euclidean metrics, that properly represent the adjacency relationships among the observations over the domain. As an insightful illustration of the potential of the methodology, we consider the spatial prediction of density data in non-convex and irregularly shaped estuarine system.

E1209: **On the compositional interpretation of frequency data**
Presenter: **Tamas Rudas**, Hungarian Academy of Sciences Centre for Social Sciences, Hungary
Co-authors: Anna Klimova
The aim is to discuss the feasibility of the standard normalization step in the compositional interpretation of frequency data, relying on some recent results in categorical data analysis. In the latter setup, the target of the statistical inference is the probability distribution underlying the data. Recent results indicate that depending on the model one has for the underlying probability distribution, the normalization may not be feasible. In the case of compositional data, the assumed distribution for the compositions is parallel to the model for the probability distribution in the case of categorical data analysis. In cases, when normalization is not possible, a proper analysis will be suggested and illustrated with data from cell biology.

E0739: **A new method for variable selection in a two and multi-group case**
Presenter: **Jan Walach**, TU Wien, Austria
Co-authors: Peter Filzmoser, Karel Hron, Beata Walczak, Lukas Najdekr
One of the main goals in metabolomics is the identification of diagnostically important variables that allow to distinguish between different patient groups. Because of the so-called size-effect, which occurs due to different concentration, conventional variable selection methods cannot be directly applied on measured data. Rather, it is necessary to make the measurements for the different samples comparable, which is possible by using specific data transformations. An alternative is to investigate the so-called relative information, which consists of the relations (ratios) between the different variables. These are independent of the size-effect and can directly be used for data processing. Relative information is analyzed by the so-called log-ratio approach, which is a standard approach in the analysis of compositional data which follows a geometrical concept endowed with the Euclidean space structure. For the purpose of variable selection, the log-ratios between all pairs of variables are employed, and they are computed separately for the different patient groups. Potential marker variables are supposed to show different variability of the pairwise log-ratios in the different groups. The variability can be estimated robustly in order to downweight the influence of data artifacts. This method turns out to have clear advantages over traditional variable selection methods in this context, in particular for problems with different group sizes, and in presence of data outliers.

EO488 Room MAL B34 RECENT DEVELOPMENTS FOR FUNCTIONAL DATA EXPLORATION Chair: Frederic Ferraty

E0349: **Model-based co-clustering for functional data**
Presenter: **Julien Jacques**, Universite de Lyon, France
As a consequence of the recent policies for smart meter development, electricity operators are nowadays able to collect data on electricity consumption widely and with a high frequency. This is in particular the case in France where EDF will be able soon to remotely record the consumption of its 27 millions clients every 30 minutes. We propose a new co-clustering methodology, based on the functional latent block model (funLBM), which allows us to build summaries of these large consumption data through co-clustering. The funLBM model extends the usual latent block model to the functional case by assuming that the curves of one block live in a low-dimensional functional subspace. Thus, funLBM is able to model and cluster large data set with high-frequency curves. An SEM-Gibbs algorithm is proposed for model inference. An ICL criterion is also derived to address the problem of choosing the number of row and column groups. Numerical experiments on simulated and original Linky data show the usefulness of the proposed methodology.

E0342: **Data depth for discontinuous functions and random sets**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic

Data depth is a mapping, which to a point in a multivariate vector space $s \in S$ and a probability measure P on S assigns a number $D(s;P)$ describing how central s is with respect to P, in an attempt to generalize quantiles to multivariate data. For S having infinite dimension, depth is typically considered only for S being the Banach space of continuous functions over a compact interval. We explore possibilities of extension of known depth functionals beyond this simplest setting. We discuss definitions, theoretical properties, and consistency/measurability issues connected with a straightforward generalization of commonly used depth functionals towards multidimensional random mappings which may lack continuity, be observed discretely, or be contaminated with additive noise.

E0730: **An asymptotic factorization of the small-ball probability: Theory and applications**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Jean-Baptiste Aubin, Aldo Goia

Recent results on an asymptotic factorization of the Small–Ball Probability (for a separable-Hilbert valued process), as the radius of the ball tends to zero are reviewed. This factorization involves a volumetric term, a pseudo-density for the probability law of the process, and a correction factor. Estimators of the latter two factors are introduced, some of their theoretical and practical aspects are illustrated.

E1109: **Operator-valued kernels for learning from functional responses**
Presenter: **Hachem Kadri**, Aix-Marseille University, France

The problems of supervised classification and regression are considered in the case where attributes and labels are functions: a data is represented by a set of functions, and the label is also a function. We focus on the use of reproducing kernel Hilbert space theory to learn from such functional data. Basic concepts and properties of kernel-based learning are extended to include the estimation of function-valued functions. In this setting,the representer theorem is restated, a set of rigorously defined infinite-dimensional operator-valued kernels that can be valuably applied when the data are functions is described, and a learning algorithm for nonlinear functional data analysis is introduced. The methodology is illustrated through speech and audio signal processing experiments.

EO394 Room MAL B35 HIGH-DIMENSIONAL INFERENCE Chair: Rajen Shah

E1259: **Significance testing in non-sparse high-dimensional linear models**
Presenter: **Jelena Bradic**, University of California San Diego, United States

In high-dimensional linear models the sparsity assumption is typically made, stating that most of the parameters are equal to zero. Under the sparsity assumption, estimation and, recently, inference have been well studied. However, in practice, sparsity assumption is not checkable and more importantly is often violated, with a large number of covariates expected to be associated with the response, indicating that possibly all, rather than just a few, parameters are non-zero. A natural example is a genome-wide gene expression profiling, where all genes are believed to affect a common disease marker. We show that existing inferential methods are sensitive to the sparsity assumption, and may, in turn, result in the severe lack of control of Type-I error. We propose a new inferential method, named CorrT, which is robust and adaptive to the sparsity assumption. CorrT is shown to have Type I error approaching the nominal level and Type II error approaching zero, regardless of how sparse or dense the model. In fact, CorrT is also shown to be optimal whenever sparsity holds. Numerical and real data experiments show a favorable performance of the CorrT test compared to the state-of-the-art methods.

E1057: **Network modeling of high-dimensional time series in the presence of factors**
Presenter: **Sumanta Basu**, Cornell University, United States

Estimating network connectivity among multiple time series is an important problem in many economic and financial applications. Examples include macroeconomic forecasting, system-wide risk monitoring and portfolio optimization. Sparsity-inducing regularizers commonly used in high-dimensional statistics do not account for the presence of pervasive factors influencing the underlying dynamics. We address the problem of estimating the networks of intertemporal and contemporaneous connectivity among multiple time series in the presence of common factor processes. For models with observable factor processes, we propose a regularized maximum likelihood procedure to estimate the factor loadings and conditional independence structure among the idiosyncratic components. In the presence of latent factors, we propose a low-rank and sparse modeling strategy to estimate the network after accounting for the underlying common factors. We demonstrate the advantage of the proposed methods on numerical experiments and a motivating application from financial economics.

E0904: **A feature distributed framework for large-scale sparse regression**
Presenter: **Chenlei Leng**, Warwick, United Kingdom

Large-scale data with a large number of features are increasingly encountered. A framework is presented for sparse high-dimensional linear regression by distributing features to multiple machines. Our method performs similarly to the infeasible oracle estimator in a centralized setting for which all the data are fitted on a single machine. Remarkably, this performance is achieved for elliptically distributed features including Gaussian variables as a special case, for any heavy tailed noises with a finite second moment, for sparse and weakly sparse signals, and for most popular sparse regression methods. Rather surprisingly, we show that a lower bound of the convergence rate of the resulting estimator does NOT depend on the number of machines. Extensive numerical studies are presented to illustrate its competitive performance.

E0801: **The xyz algorithm for fast interaction search in high-dimensional data**
Presenter: **Gian-Andrea Thanei**, ETH Zurich, Switzerland
Co-authors: Rajen Shah, Nicolai Meinshausen

When performing regression on a dataset with p variables, it is often of interest to go beyond using main linear effects and include interactions as products between individual variables. For small-scale problems, these interactions can be computed explicitly but this leads to a computational complexity of at least $O(p^2)$ if done naively. This cost can be prohibitive if p is very large. We introduce a new randomised algorithm that is able to discover interactions with high probability and under mild conditions has a runtime that is subquadratic in p. We show that strong interactions can be discovered in almost linear time, whilst finding weaker interactions requires $O(p^a)$ operations for $1 < a < 2$ depending on their strength. The underlying idea is to transform interaction search into a closest pair problem which can be solved efficiently in subquadratic time. The algorithm is called xyz and is implemented in the language R. We demonstrate its efficiency for application to genome-wide association studies, where more than 10^11 interactions can be screened in under 280 seconds with a single-core 1.2 GHz CPU.

EO236 Room G21A REGRESSION MODELS UNDER NON I.I.D. SETTINGS Chair: Sophie Dabo

E0208: **Some nonparametric tests for change-point detection based on the P-P and Q-Q plot processes**
Presenter: **Salim Bouzebda**, Universite de Technologie de Compiegne, France
Nonparametric procedures are proposed for testing change-point by using the P-P and Q-Q plots processes. The limiting distributions of the proposed statistics are characterized under the null hypothesis of no change and also under contiguous alternatives. We give an estimator of the change-point coefficient and obtain its strong consistency. Emphasis is placed on the explanation of the strong approximation methodology.

E0622: **Conditional homoscedasticity test in time series with dependent innovations: Asymptotic power properties**
Presenter: **Mohamed Chaouch**, United Arab Emirates University, United Arab Emirates
On the basis of the functional limiting law of certain cumulative residual processes, we propose several tests for examining hypotheses about conditional variance functions in time series with martingale innovations. The proposed tests are asymptotically distribution-free under the null hypothesis. As an application, we are interesting to test the volatility function $\sigma^2(x)$ appearing in a non-linear autoregressive model with errors $\varepsilon_t = \sigma(X_t)e_t$, where $e_t = \rho_0 e_{t-1} + u_t$. This model provides linkage between risk and expected return of financial assets. Moreover, it can be used for testing the martingale difference sequence hypothesis, which is typically uncritically adapted in financial time series models. The limiting distribution of these tests under a local of sequence of alternatives is obtained and the asymptotic local power is derived.

E1039: **Efficient MCMC estimation for spatial econometrics models with convex combination of connectivity matrices**
Presenter: **Nicolas Debarsy**, CNRS, France
Co-authors: James LeSage
An efficient Bayesian MCMC procedure is proposed to estimate spatial econometrics models in which the cross-section dependence scheme is modeled through a convex combination of connectivity structures, each one representing a motivation for interactions between observations. An MCMC approach has been previously proposed to estimate and perform inference on the parameters of the convex combination in these models. Even though such an approach works well when the number of matrices included in the convex combination is low (2 or 3), the computer time required to estimate these models increases a lot when more matrices are considered. This comes from the need to pre-compute the Jacobian of the transformation of these models for a grid of values for all parameters of the convex combination. We propose an efficient MCMC approach to estimate these models. It is constructed from block sampling based on reversible jump MCMC and relies on bounded inference MCMC based on the lower and upper bounded conditional distributions. We show that our approach is not affected by the number of considered matrices in the convex combination, nor by the number of draws performed.

E1603: **A nonstationarity copula-based conditional quantile approach: Application to extreme daily stream-flow in Canada**
Presenter: **Bouchra Nasri**, HEC Montreal, Canada
Co-authors: Bruno Remillard, Taoufik Bouezmarni
Hydrological frequency analysis is commonly used by engineers and hydrologists to provide the basic information on planning, design, and management of hydraulic and water resources systems under the assumption of stationarity. However, with increasing evidence of climate change, the assumption of stationarity, a prerequisite for traditional frequency analysis, might be invalid, making the conventional analysis pointless. Recently, some approaches were proposed in order to estimate extreme conditional quantiles in nonstationary frequency analysis settings, namely the so-called covariate method, where the covariates are incorporated into the parameters of the distribution function of the response variable, and the quantile regression method. However, these two methods cannot fully describe the dependence between the variable of interest and its covariates. In order to overcome this limitation, in addition to dealing with nonstationarity, we propose a new estimator for conditional quantile based on copula functions. We study its asymptotic behavior and we suggest a bootstrap procedure in order to construct uniform confidence bands around the conditional quantile function; as a by-product, we also obtain a formal goodness-of-fit test. Finally, we present a simulation study demonstrating the finite sample performance of the proposed estimator and we illustrate its usefulness with an application to a hydro-climatic dataset.

EO134 Room CLO 101 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING I Chair: Raffaele Argiento

E0689: **Using particle Gibbs methods with Bayesian nonparametric models**
Presenter: **Jim Griffin**, University of Kent, United Kingdom
Bayesian nonparametric models are characterized by an infinite number of parameters. In many nonparametric models, only a finite number of these parameters are used for a finite sample but the exact number is unknown. For example, with a finite sample, a Dirichlet process mixture model will only have a finite (but unknown) number of clusters or a Indian buffet process factor model will only have a finite (but unknown) number of factors. Defining effective Markov chain Monte Carlo (MCMC) methods for posterior inference in these models is challenging and often uses one-at-a-time updates which can lead to slow mixing. Particle MCMC methods are attractive as they can avoid one-at-a-time updates but defining effective particle filters for these models is also very challenging. We will consider versions of particle Gibbs methods which are useful with these models and certain time-dependent extensions (where particle filtering is a natural method for inference). The methods will be illustrated on a range of Bayesian nonparametric models and data settings.

E1275: **Geometric MCMC for infinite-dimensional inverse problems**
Presenter: **Alexandros Beskos**, University College London, United Kingdom
Bayesian inverse problems often involve sampling posteriors on infinite-dimensional spaces. Traditional MCMC algorithms are characterized by deteriorating mixing times upon mesh-refinement, when finite-dimensional approximations become more accurate. Such methods are typically forced to reduce step-size as the discretization gets finer, thus are expensive as a function of dimension. Recently, a new class of MCMC methods with mesh-independent convergence times has emerged. However, few of them take into account the geometry of the posterior informed by the data. At the same time, recently developed geometric MCMC algorithms are found to be powerful in exploring complicated distributions that deviate from elliptic Gaussian laws, but are in general computationally intractable for models in infinite dimensions. We combine geometric methods on a finite-dimensional subspace with mesh-independent infinite-dimensional approaches. Our objective is to speed up MCMC mixing times, without significantly increasing computational costs. This is achieved by using ideas from geometric MCMC to probe the complex structure of an intrinsic finite-dimensional subspace where most data information concentrates, while retaining robust mixing times as the dimension grows by using pCN-like methods in the complementary subspace. The resulting algorithms are demonstrated in the context of 3 challenging inverse problems arising in subsurface flow, heat conduction and incompressible flow control.

E1242: **Design of informed local proposals for MCMC in discrete spaces**
Presenter: **Giacomo Zanella**, Bocconi University, Italy
There is a lack of methodological results to guide practitioners in designing efficient Markov chain Monte Carlo (MCMC) algorithms in discrete spaces. For example, it is still unclear how to extend gradient-based MCMC (e.g. Langevin and Hamiltonian schemes) to networks or partitions spaces. This is particularly relevant when fitting Bayesian nonparametric models, which often involve combinatorial and discrete latent structures.

Motivated by this observation, we consider the problem of designing appropriate informed MCMC proposal distributions in discrete spaces. In particular: assuming perfect knowledge of the target measure, what is the optimal Metropolis-Hastings proposal given a fixed set of allowed local moves? Under regularity assumptions on the target, we derive the class of asymptotically optimal proposal distributions, which we call locally-balanced Proposals. Such proposals are maximal elements, in terms of Peskun ordering, among proposals obtained as pointwise transformations of the target density. This class of proposals includes Langevin MCMC as a special case and can be seen as a generalization of gradient-based methods to discrete frameworks. We discuss asymptotic analysis, applications to discrete frameworks and connections to other schemes (e.g Multiple-Try).

E0381: **Inferring components and clusters in Bayesian finite mixture modelling**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Business and Economics, Austria
Co-authors: Sylvia Fruhwirth-Schnatter, Bettina Gruen

The selection of a suitable number of mixture components is a difficult problem in finite mixture modelling without a generally accepted solution so far. The choice of the number of mixture components is in general crucial, because it is assumed that each mixture component captures exactly one data cluster. In the Bayesian framework, a natural approach to estimate the number of components is to treat it as an unknown model parameter, specify a prior on it and determine its posterior distribution. Several inference methods have been proposed for estimating this distribution, in particular, reversible jump Markov chain Monte Carlo (RJMCMC) techniques. However, it can be a difficult task to design suitable proposal densities in higher-dimensional parameter spaces for these samplers. Based on recently published results, we propose an alternative approach where we distinguish between the total number of mixture components in the mixture model and the number of 'active' components, i.e. components, to which observations are assigned and which correspond to the data clusters. In this way we are able to make inference on both the number of mixture components and the number of data clusters without employing complex transdimensional sampling techniques, but rather using only familiar Gibbs sampling techniques. This approach is illustrated using simulation studies and real applications.

EO581 Room CLO 102 INFERENCE FOR DIFFUSION MODELS Chair: Frank van der Meulen

E0346: **Nonparametric learning of stochastic differential equations**
Presenter: **Andreas Ruttor**, TU Berlin, Germany
Co-authors: Philipp Batz, Manfred Opper

Differential equations are suitable models for many real-world systems, but prior knowledge about their structure may be limited or missing. While parametric drift and diffusion functions could be chosen by model selection, it is often difficult to specify a good set of candidates in order to avoid model mismatch. A better approach is to use nonparametric Bayesian inference for estimating complete functions. That way both drift and diffusion functions for a system of stochastic differential equations are learned from observations of the state vector. Gaussian processes are used as flexible models for these functions and estimates are calculated directly from dense data sets using Gaussian process regression. In case of sparse observations the latent dynamics between data points is estimated by an approximate expectation maximisation algorithm. Here the posterior over states is approximated by a piecewise linearized process of the Ornstein-Uhlenbeck type and the maximum a posteriori estimation of the drift is facilitated by a sparse Gaussian process approximation.

E0930: **Correlated pseudo marginal schemes for partially observed diffusion processes**
Presenter: **Andrew Golightly**, Newcastle University, United Kingdom

Stochastic differential equations (SDEs) provide a natural framework for modelling intrinsic stochasticity inherent in many continuous-time physical processes. Performing fully Bayesian inference for such models, using discrete time data that may be incomplete and subject to measurement error is a challenging problem. One widely used approach is to replace unavailable transition densities with an Euler-Maruyama approximation, made sufficiently accurate through the introduction of intermediate times between observations. Pseudo marginal MCMC schemes are increasingly used for the inference problem, since for a given discretisation level, the observed data likelihood can be unbiasedly estimated using a particle filter. When observations are particularly informative, a diffusion bridge construct can be used to drive the particle filter. Recent work in state-space settings has shown how the pseudo marginal approach can be made much more efficient by correlating the underlying pseudo random numbers used to form the estimate of likelihood at the current and proposed values of the unknown parameters. We extend this approach to the discretised diffusion process framework by correlating the Gaussian innovations that drive the diffusion bridge construct used inside particle filter. We find that the resulting approach offers substantial gains in efficiency over a standard implementation.

E1180: **Inference for diffusion processes from observations of passage times**
Presenter: **Moritz Schauer**, Leiden University, Netherlands

A diffusion process with unknown parameters is indirectly observed with observations being a sequence of passage times. We introduce a method to accurately simulate the conditional diffusion process given the observed random times. A change of measure is applied to a proposal process with a guiding drift term derived from an approximation of the conditional process. The method is used in a Markov chain Monte Carlo procedure to sample from the joint posterior distribution of the unobserved diffusion trajectory and the model parameters given the observed random times. This is illustrated fitting a diffusion model for neuronal spike generation to observations of spike times.

E1163: **MCMC inference for discretely-observed diffusions: Improving efficiency**
Presenter: **Christiane Fuchs**, Helmholtz Center Munich, Germany
Co-authors: Susanne Pieschner

Diffusion processes are used to realistically model time-continuous processes in biology. In real-data applications, inference for diffusions is difficult when the likelihood function is analytically unavailable. A widely applicable method for parameter estimation is a Markov chain Monte Carlo (MCMC) approach which simulates and employs synthetic sample paths. For high-dimensional processes, however, the method is computationally intensive. We present attempts to improve estimation efficiency in practice, including approaches that did not turn out to be successful. We illustrate the inferential methods on time-resolved single-cell gene expression data.

EO642 Room Senate HIGH DIMENSIONAL TIME SERIES MODELS AND THEIR APPLICATIONS Chair: George Michailidis

E0919: **Regularized estimation and testing for high-dimensional multi-block VAR models**
Presenter: **Jiahe Lin**, University of Michigan, United States
Co-authors: George Michailidis

Dynamical systems comprising of multiple components that can be partitioned into distinct blocks originate in many scientific areas. A pertinent example is the interactions between financial assets and selected macroeconomic indicators, e.g. a stock index and an employment index. A key shortcoming of this approach is that it ignores potential influences from other related components that may exert influence on the system's dynamics. To mitigate this issue, we consider a multi-block linear dynamical system with Granger-causal ordering between blocks. We derive the MLE for the posited model for Gaussian data in the high-dimensional setting based on appropriate regularization schemes. To optimize the underlying non-convex likelihood function, we develop an iterative algorithm with convergence guarantees. We establish theoretical properties of the MLE, leveraging the decomposability of the regularizers and a careful analysis of the iterates. Finally, we develop testing procedures for

the null hypothesis of whether a block "Granger-causes" another block of variables. The performance of the model and the testing procedures are evaluated on synthetic data, and illustrated on a data set involving log-returns of the US S&P 100 component stocks and key macroeconomic variables.

E1430: **Gaussian approximation for high dimensional time series**
Presenter: **Danna Zhang**, University of California, San Diego, United States
Co-authors: Wei Biao Wu
High-dimensional time series data arise in a wide range of disciplines. The fact that the classical CLT for i.i.d. random vectors may fail in high dimensions makes high-dimensional inference notoriously difficult. More challenges are imposed by temporal and cross-sectional dependence. We will introduce the high-dimensional CLT for time series. Its validity depends on the sample size n, the dimension p, the moment condition and the dependence of the underlying processes. An example is taken to appreciate the optimality of the allowed dimension p. Equipped with the high-dimensional CLT result, we have a new sight on many problems such as inference for covariances of high-dimensional time series which can be applied in the analysis of network connectivity, inference for multiple posterior means in MCMC experiments, Kolmogorov-Smirnov test for high-dimensional time series data as well as inference for high-dimensional VAR models. We will also introduce an estimator for long-run covariance matrices and two resampling methods, i.e., Gaussian multiplier resampling and subsampling, to make the high-dimensional CLT more applicable. The results are then corroborated by a simulation study of an MCMC experiment based on a hierarchical model and real data analysis for high-dimensional time series.

E1524: **Autoregressive model for matrix-valued time series**
Presenter: **Han Xiao**, Rutgers University, United States
In finance, economics and many other fields, observations in a matrix form are often observed over time. For example, several key economic indicators are reported in different countries every quarter. Various financial characteristics of many companies are reported over time. Import-export figures among a group of countries can also be structured in a matrix form. Although it is natural to turn the matrix observations into a long vector then use standard vector time series models, it is often the case that the columns and rows of a matrix represent different sets of information that are closely interplayed. We propose a novel matrix autoregressive model that maintains and utilizes the matrix structure to achieve greater dimensional reduction as well as easier interpretable results. The model can be further simplified by a set of reduced rank assumptions. Estimation procedure and its theoretical properties are investigated and demonstrated with simulated and real examples.

E1570: **A system-wide approach to measure connectivity in the financial sector**
Presenter: **Sreyoshi Das**, Cornell University, United States
Co-authors: Sumanta Basu, George Michailidis, Amiyatosh Purnanandam
The aim is to introduce and estimate a model that leverages a system-wide approach to identify systemically important financial institutions. It is based on a recently developed Lasso penalized Vector Auto-regressive (LVAR) model, that exhibits desirable statistical properties and enables us to detect systemic events and key institutions associated with them. The model explicitly allows for the possibility of connectivity amongst all institutions under consideration: this is in sharp contrast with extant measures of systemic risk that, either explicitly or implicitly, estimate such connections using pair-wise relationships between institutions. Using simulations we show that our approach can provide considerable improvement over extant measures in detecting systemically important institutions. Finally, we estimate our model for large financial institutions in the U.S. and show its usefulness in detecting systemically stressful periods and institution with real data.

EO238 Room CLO 203 STATISTICAL SIZE DISTRIBUTIONS AND HEAVY-TAILED DISTRIBUTIONS Chair: Christophe Ley

E0544: **κ-generalized models of income and wealth distributions: A survey**
Presenter: **Fabio Clementi**, University of Macerata, Italy
Co-authors: Mauro Gallegati
A survey of results related to the "κ-generalized distribution" is provided. κ-generalized distribution is a statistical model for the size distribution of income and wealth. Topics include, among others, discussion of basic analytical properties, interrelations with other statistical distributions as well as aspects that are of special interest in the income distribution field, such as the Gini index and the Lorenz curve. An extension of the basic model that is most able to accommodate the special features of wealth data is also reviewed. The given survey of empirical applications shows the κ-generalized models of income and wealth to be in excellent agreement with the observed data in many cases.

E0539: **Pareto type probability distribution for cylindrical data**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan
Co-authors: Kunio Shimizu, Toshihiro Abe
A probability distribution on the cylinder, or cylindrical distribution provides a bivariate distribution with linear and circular random variables. We propose a cylindrical distribution which has a heavy-tailed linear marginal distribution. Its conditional distribution of the linear variable given circular variable is a generalized Pareto distribution, so it might not have any conditional moments. However, since the conditional mode and median are expressed by the closed forms, we can use these indices to represent the relation between the linear and circular variables. The circular marginal distribution is a wrapped Cauchy distribution and conditional distribution of the circular variable given linear variable is the family of Jones and Pewsey distribution. These properties leads to the application for the cylindrical data whose linear variable might have large values and circular variable is symmetric. As an example of applications, we fit the proposed distribution to the seismic magnitude and sequences of epicenter data and compare the result with that by other cylindrical distribution which has a light-tailed linear marginal distribution.

E0841: **Characterizing the concentration of tail risk under fat-tails**
Presenter: **Pasquale Cirillo**, Delft University of Technology, Netherlands
The aim is to discuss how concentration measures can be used to study fat-tails, when dealing with the size distributions of income, wealth, losses, etc. Focusing on the Gini index, we will go beyond its use as a concentration measure and will show how to define it as a probabilistic measure of risk, and in particular of tail risk, deriving a completely brand new set of tools for the study of fat-tails. The goal is to better characterize tail risk in the case of a very large or infinite population variance. We will naturally explain the main theoretical aspects and properties, but we will also discuss heuristics and applications on actual data.

E0744: **Size distributions and the moment problem**
Presenter: **Christian Kleiber**, Universitaet Basel, Switzerland
The moment problem asks whether or not a given distribution is uniquely determined by its moments. The existence of moment-indeterminate distributions has been known since the late 19th century; however, first examples were often considered as mere mathematical curiosities. Recent research has shown that the phenomenon is more widespread than previously thought. Specifically, distributions that are not determined by their moments arise in the modelling of size distributions. We study determinacy issues for selected size distributions occurring in economics and actuarial science.

EO049 Room CLO 204 DESIGN, MODELING, DATA ANALYSIS, AND COMPUTATIONAL STATISTICS Chair: Subir Ghosh

E0475: **Designing an adaptive trial with treatment selection and a survival endpoint**
Presenter: **Christopher Jennison**, University of Bath, United Kingdom
A clinical trial is considered in which two versions of a new treatment are compared against control with the primary endpoint of overall survival. At an interim analysis, mid-way through the trial, one of the two treatments is selected, based on the short term response of progression free survival. For such an adaptive design the family-wise type I error rate can be protected by use of a closed testing procedure to deal with the two null hypotheses and combination tests to combine data from before and after the interim analysis. However, with the primary endpoint of overall survival, there is still a danger of inflating the type I error rate: we present a way of applying the combination test that solves this problem simply and effectively. With the methodology in place, we then assess the potential benefits of treatment selection in this adaptive trial design.

E0742: **Modelling multilevel data under complex sampling designs: An empirical likelihood approach**
Presenter: **Yves Berger**, University of Southampton, United Kingdom
Co-authors: Melike Oguz-Alper
Data used in social, behavioural, health or biological sciences may have a hierarchical structure due to the population of interest or the sampling design. Multilevel are often used to analyse such hierarchical data, or to estimate small domains means. These data are often selected with unequal probabilities from a clustered and stratified population. Inference may be invalid when the sampling design is informative. We propose a design-based empirical likelihood approach for the regression parameters of a multilevel model. It has the advantage of taking into account of the informative effect of the sampling design. This approach can be used for point estimation, hypothesis testing and confidence intervals for the sub-vector of parameters. It can be also used for estimation of small domains means. It provides asymptotically valid inference for the finite population parameters. The simulation study shows the advantages of the empirical likelihood approach over alternative approaches used for multilevel models. We applied our approach to the Programme for International Student Assessment (PISA) survey data. We show that the proposed approach can give more accurate and different p-values than the naive restricted maximum likelihood approach, which ignores the survey weights.

E0471: **Construction of two-level factorial and fractional factorial designs with runs in blocks of size two**
Presenter: **Janet Godolphin**, Surrey, United Kingdom
In many experiments involving factorial and fractional factorial designs, attention is focused on estimation of all main effects and two factor interactions. Design construction are considered when, due to practical constraints, runs are arranged in blocks of size two. For p factors, and M at least as large as a given function of p, a construction approach is provided which generates all designs in which M replicates are arranged in blocks of size two so that all main effects and two factor interactions are estimable. The method incorporates recognition of isomorphic designs to avoid double counting. A design ranking is proposed to give guidance on design selection which prioritises estimation of main effects. This is useful in practice since for some p, M combinations the number of designs is large (for example, for $p = 8$ and $M = 4$ there are 343 designs) and there can be considerable variation in the quality of estimation between designs. The full factorial designs can be used as a source of root designs for construction of designs in fractional replicates, again in blocks of size two. The method is illustrated by examples with up to $p = 15$ factors.

E0456: **CV, ECV, and robust CV designs for replications under a class of linear models in factorial experiments**
Presenter: **Subir Ghosh**, University of California, United States
A class of linear models is considered for describing the data collected from an experiment. Any two models have some common as well as uncommon parameters. A common variance (CV) design is proposed for collecting the data so that all the uncommon parameters are estimated with as similar variances as possible in all models. The variance equality for a CV design is attained exactly when there is one uncommon parameter for any two models within the class. A new concept robust CV designs for replications having the possibility of replicated observations is introduced. The conditions are presented for a CV design having no replicated observations to be robust for general replicated observations. A CV design having no replicated observations is always robust for any equally replicated observations. Examples of Efficient CV (ECV) designs as well as robust CV designs for general replicated observations are also presented. A simple illustrative example of the complete 2 x 3 factorial design is demonstrated to be not a CV design and then the condition on replications of each run is obtained to turn it into a CV design.

EO421 Room MAL 402 ADVANCES IN ORDINAL DATA ANALYSIS Chair: Cristina Mollica

E0323: **The use of penalized likelihood for analyzing ranking data**
Presenter: **Mayer Alvo**, University of Ottawa, Canada
Co-authors: Hang Xu
Different score functions are considered in order summarize certain characteristics for one and two sample ranking data sets. Our approach is flexible and is based on embedding the nonparametric problem in a parametric framework. We make use of the von Mises-Fisher distribution to approximate the normalizing constant in our model. In order to gain further insight in the data, we make use of penalized likelihood to narrow down the number of items where the rankers differ. We applied our method on various real life data sets and we conclude that our methodology is consistent with the data.

E0561: **Probabilistic preference learning with the Mallows rank model**
Presenter: **Marta Crispino**, Bocconi, Italy
Co-authors: Valeria Vitelli, Arnoldo Frigessi, Oystein Sorensen, Elja Arjas
Ranking and comparing items is crucial for collecting information about preferences in many areas, from marketing to politics. The Mallows rank model is among the most successful approaches to analyze rank data, but its computational complexity has limited its use to a particular form based on Kendall distance. We develop new computationally tractable methods for Bayesian inference in Mallows models that work with any right-invariant distance. Our method performs inference on the consensus ranking of the items, also when based on partial rankings, such as top-k items or pairwise comparisons. We prove that items that none of the assessors has ranked do not influence the maximum a posteriori consensus ranking, and can therefore be ignored. When assessors are many or heterogeneous, we propose a mixture model for clustering them in homogeneous subgroups, with cluster-specific consensus rankings. We develop approximate stochastic algorithms that allow a fully probabilistic analysis, leading to coherent quantifications of uncertainties. We make probabilistic predictions on the class membership of assessors based on their ranking of just some items, and predict missing individual preferences, as needed in recommendation systems. We test our approach using several experimental and benchmark datasets.

E0528: **Modelling multiple Likert items through an adjacent categories ordinal paired comparison model**
Presenter: **Brian Francis**, Lancaster University, United Kingdom
Co-authors: Regina Dittrich
The modelling of multiple Likert items measured on the same Likert scale can be challenging. We review existing methodologies, which can be classified into two approaches - joint modelling of the absolute Likert ratings, and modelling of the relative differences between Likert Items. One

approach for modelling relative differences has been suggested, which led to a paired comparison model with three outcomes for each pair of Likert items the first item preferred, the second item preferred and no preference. Such a model can be fitted as a Poisson log-linear model, with each item placed on a "worth" (0,1) scale; in addition, covariates can be included. That model is extended to allow for a distance metric between the pairs of items. This leads to an adjacent categories ordinal paired comparison model with ties. This model makes better use of the underlying Likert rating scale, and we compare the two approaches through worth plots. Numerical issues and implementation in R are both discussed, together with typical restrictions on the size of the problem that can be dealt with. Future work will be discussed.

E0872: **Assessment of zero inflated mixture models for ordinal data**
Presenter: **Maria Iannario**, University of Naples Federico II, Italy
Co-authors: Rosaria Simone
Excess of zeros is a commonly encountered phenomenon that limits the use of traditional regression models for analysing ordinal data which exhibit zero inflation. These data concern contexts where respondents express a graduated perception on a specific item or experiments identify levels of increasing assessments including zero (subjects are not susceptible to express the response). Specifically, the zero counts could be simply absent in the rating of respondents for the absence of the symptom or activity (structural zeros) or present with low frequency not observed because of sampling variation (sampling zeros). The focus is on modelling the data by means of zero-inflated mixture models by taking into account both excess of zeros and heterogeneity. It has been designed so as to discriminate between structured and unstructured zeros by placing particular emphasis on the uncertainty concerning the evaluation process. The performance of the proposed model is assessed through simulation studies and survey data.

EO384 Room MAL 414 STATISTICS IN BIOSCIENCE Chair: Hao Chen

E1303: **Approximate confidence distribution computing (ACC): A likelihood-free method with statistical guarantees**
Presenter: **Min-ge Xie**, Rutgers University, United States
Approximate Bayesian computing (ABC) is a likelihood-free method that has grown increasingly popular since early applications in population genetics. The purpose is to consider the use of ABC method in frequentist application and, in particular, its extended version based on the concept of confidence distribution. The extended version, called approximate confidence distribution computing (ACC), can overcome two defects of the traditional ABC method, namely, lack of theory supporting the use of non-sufficient summary statistics and lack of guardian for the selection of prior. It is also demonstrated that a well-tended ACC algorithm can greatly increase its computing efficiency over a traditional ABC algorithm. Related theories on frequentist coverage, both asymptotic and exact, are investigated. The method is also illustrated with both simulations and real data example of tuberculosis transmission.

E0277: **Two-phase designs for joint trait- and genotype-dependent sampling in post-GWAS regional sequencing**
Presenter: **Radu Craiu**, University of Toronto, Canada
Co-authors: Shelley Bull, Osvaldo Espin-Garcia
In the post-GWAS era, identifying causal variants and susceptibility genes in GWAS-identified regions of association has become an important goal for researchers. Despite decreasing costs of next-generation sequencing (NGS) technologies, sequencing all subjects in large-scale studies is still prohibitive. Two-phase designs are proposed when findings from genome-wide association study (GWAS) are followed by regional sequencing that is too expensive to implement in the entire cohort. Inference is based on a semiparametric likelihood that is optimized using the EM algorithm. A GWAS-SNP serves as a surrogate covariate in inferring association between a sequence variant and a normally distributed quantitative trait (QT). Simulations are used to quantify the efficiency and power of joint QT-SNP-dependent sampling and analysis under alternative sample allocations. Joint allocation balanced on SNP genotype and extreme-QT strata yields significant power improvements compared to marginal QT- or SNP-based allocations. A sequencing study of systolic blood pressure is used to illustrate the method.

E1182: **Estimation and testing of survival functions via generalized fiducial inference with censored data**
Presenter: **Jan Hannig**, University of North Carolina at Chapel Hill, United States
Co-authors: Yifan Cui
Fiducial Inference has a long history, which at times aroused passionate disagreements. However, its application has been largely confined to relatively simple parametric problems. We present what might be the first time fiducial inference is systematically applied to estimation of a nonparametric survival function under right censoring. We find that the resulting fiducial distribution gives rise to surprisingly good statistical procedures applicable to both one sample and two sample problems. We establish a functional Bernstein-von Mises theorem, and perform thorough simulation studies in scenarios with different levels of censoring. The proposed fiducial based confidence intervals maintain coverage in situations where asymptotic methods often have substantial coverage problems. Furthermore, the average length of the proposed confidence intervals is often shorter than the length of competing methods that maintain coverage. Finally, the proposed fiducial test is more powerful than various types of log-rank tests and sup log-rank tests in some scenarios. We illustrate the proposed fiducial test comparing chemotherapy against chemotherapy combined with radiotherapy using data from the treatment of locally unresectable gastric cancer.

E0860: **Transformed dynamic quantile regression on censored data**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong
A class of power-transformed linear quantile regression models is proposed for time- to-event observations that are subject to conditionally independent censoring. By introducing power transformation with different transformation parameters for individual quantile levels, the proposed model relaxes the limitation due to the global linear assumption required in a previous formulation, and thus, it enjoys a greater extent of flexibility. Our framework provides simultaneous estimation of various quantiles and is amongst the first which considers power transformation as a process with respect to the quantile values. The uniform consistency and weak convergence of the proposed estimator as a process with respect to a sequence of quantile levels are established. Simulation studies and applications to real data sets are presented for verifying the performance of the proposed estimator our empirical results is shown to outperform existing contenders under various settings.

EO439 Room MAL 415 SIMULTANEOUS STATISTICAL INFERENCE Chair: Thorsten Dickhaus

E0525: **Post-selection inference via multiple testing**
Presenter: **Pierre Neuvial**, CNRS and Toulouse Mathematics Institute, France
Co-authors: Gilles Blanchard, Etienne Roquain
The objective is to devise multiple testing procedures able to provide a statistical guarantee on any candidate set of rejected hypotheses, including user-defined and/or data-driven candidate sets. We introduce a general methodology relying on the control of a multiple testing error rate that we call the joint Family-Wise Error Rate (JER). Our construction generalizes existing so-called "post hoc" procedures under positive dependence. We propose a generic approach to build JER-controlling procedures in the situation where the joint null distribution of the test statistics is known or can be sampled from. Our theoretical statements are supported by numerical experiments.

E0535: **From post hoc analysis to post selection inference**
Presenter: **Daniel Yekutieli**, Tel-Aviv University, Israel

The aim is to explain the connection between the work of Tukey and Scheffe on post-hoc analysis, Benjamini and Hochberg's work on the FDR, the work of Efron and colleagues on the Bayesian FDR, my work with Benjamini on selective inference, the work of Berk et al. on post-selection inference, and recent work on frequentist and Bayesian post-selection inferences based on the conditional likelihood.

E0767: **All-resolution inference: Consistent confidence bounds for the false discovery proportion in high-dimensional data**
Presenter: **Jelle Goeman**, Leiden University Medical Center, Netherlands
Co-authors: Aldo Solari

The combination of closed testing with the Simes inequality can be used to derive simultaneous upper confidence bounds for the false discovery proportion (FDP) in all subsets of the hypotheses. Consequently, these bounds are valid even when these subsets are chosen post hoc. This property is useful e.g. in neuroscience, where it can be used to find interesting brain regions and to infer the percentage of activation in those brain regions, all in the same data, or in gene expression data, where it may be used to do gene set testing without requiring that the database of gene sets is chosen before seeing the data. We investigate the power properties of this method in large problems, showing that the procedure is remarkably powerful given the enormous simultaneity. We show that (1.) if sufficient signal is present, as the number of hypothesis goes to infinity, the number of discoveries allowed at any FDP-level also goes to infinity. We also show that (2.) the lower bound of the FDP seen as an estimator of FDP is uniformly consistent with respect to the size of the multiple testing problem.

E0638: **Multivariate multiple test procedures based on nonparametric copula estimation**
Presenter: **Thorsten Dickhaus**, University of Bremen, Germany
Co-authors: Andre Neumann, Taras Bodnar, Dietmar Pfeifer

Multivariate multiple test procedures have received growing attention recently. This is due to the fact that data generated by modern applications typically are high-dimensional, but possess pronounced dependencies due to the technical mechanisms involved in the experiments. Hence, it is possible and often necessary to exploit these dependencies in order to achieve reasonable power. We express dependency structures in the most general manner, namely, by means of copula functions. One class of nonparametric copula estimators is constituted by Bernstein copulae. We extend previous statistical results regarding bivariate Bernstein copulae to the multivariate case and study their impact on multiple tests. In particular, we utilize them to derive asymptotic confidence regions for the family-wise error rate (FWER) of simultaneous test procedures which are empirically calibrated by making use of Bernstein copulae approximations of the dependency structure among the test statistics. A simulation study quantifies the gain in FWER level exhaustion and, consequently, power which can be achieved by exploiting the dependencies, in comparison with common threshold calibrations like the Bonferroni or Sidak corrections. Finally, we demonstrate an application of the proposed methodology to real-life data from insurance.

EO362 Room MAL 416 STATISTICS ON NETWORKS Chair: Jean-Pierre Florens

E0479: **Identifying and estimating social connections from outcome data**
Presenter: **Aureo de Paula**, University College London, United Kingdom

Knowledge of the relevant linkages between individuals is usually necessary for the estimation of social interaction models. We obtain results that allow for the estimation of parameters of interest in a model with endogenous, exogenous and correlated effects without information on the relevant linkages. Our identification analysis relies on usual assumptions on the nature of interactions. To obtain identification we further impose conditions on the density of links and repeated observation of outcomes for a given group of individuals. We provide an estimation strategy which we investigate via simulations and an empirical illustration.

E0486: **Convolution on networks**
Presenter: **Jean-Pierre Florens**, Toulouse School of Economics, France
Co-authors: Andrii Babii

A concept of convolution on possibly discrete or continuous network is defined where the interdependence structure is characterized by a Laplacian operator. This operator determines a concept of Fourier transform and then a convolution and we concentrate our attention to linear models where the link between parameter and explanatory variable is a convolution. Regularized estimations using instrumental variables are proposed and their properties are studied. Several example are given: finite network, locally compact groups, minimum manifolds.

E1789: **Network and panel quantile effects via distribution regression**
Presenter: **Martin Weidner**, University College London, United Kingdom
Co-authors: Victor Chernozhukov, Ivan Fernandez-Val

A method is provided to construct simultaneous confidence bands for quantile functions and quantile effects in nonlinear network and panel models with unobserved two-way effects, strictly exogenous covariates, and possibly discrete outcome variables. The method is based upon projection of simultaneous confidence bands for distribution functions constructed from fixed effects distribution regression estimators. These fixed effects estimators are bias corrected to deal with the incidental parameter problem. Under asymptotic sequences where both dimensions of the data set grow at the same rate, the confidence bands for the quantile functions and effects have correct joint coverage in large samples. An empirical application to gravity models of trade illustrates the applicability of the methods to network data.

E1798: **Inference on social effects when the network is sparse and unknown**
Presenter: **Eric Gautier**, Toulouse School of Economics, France
Co-authors: Christiern Rose

Models of social interaction are considered when the underlying networks are unobserved but sparse and there are endogenous, contextual, and correlated effects. We accommodate prior knowledge on the sparsity pattern (group sparsity, known existing or nonexisting links) and restrictions on the parameters. We provide results on identification, rates of convergence, model selection, and inference for the parameters and linear functionals in the high-dimensional paradigm. The inference is robust to identification and uniform over large classes of sparse identifiable parameters and data generating processes. Some results hold in finite samples.

EO286 Room MAL 421 MACHINE LEARNING, APPROXIMATION, AND ROBUSTNESS Chair: Andreas Christmann

E1779: **Regression phalanxes**
Presenter: **Ruben Zamar**, University of British Columbia, Canada
The notion of phalanxes has been recently introduced in the context of rare-class detection in two-class classification problems. A phalanx is a subset of features that work well for classification tasks. We propose a different class of phalanxes for application in regression settings. We define a *regression phalanx* as a subset of features that work well together for prediction. We propose a novel algorithm which automatically chooses regression phalanxes from high-dimensional datasets using hierarchical clustering and builds a prediction model for each phalanx for further ensembling. Through extensive simulation studies and several real-life applications in various areas (including drug discovery, chemical analysis of spectra data,microarray analysis and climate projections) we show that an ensemble of regression phalanxes improves prediction accuracy when combined with effective prediction methods like lasso or random forests.

E0704: **Consistency and robustness properties of predictors based on locally learnt support vector machines**
Presenter: **Florian Dumpert**, University of Bayreuth, Germany
Among different machine learning methods, support vector machines (SVMs) play an important role in many fields of science nowadays. A lot of research about statistical and computational properties of support vector machines and related kernel methods has been done during the last two decades up to now. On the one hand, from a statistical point of view, one is interested in consistency and robustness of the method. On the other hand, from a computational point of view, one is interested in a method that can deal with many observations and many features. As SVMs need a lot of computing power and storage capacity, different ways to handle big data sets were proposed. One of them, which is called regionalization, divides the space of the declaring variables into possibly overlapping regions in a data driven way and defines the output predicting function by composing locally learnt support vector machines. It is possible to show that a predictor learnt in this way conserves consistency and robustness results under assumptions that can be checked by the user of this method. We will have a closer look at the consistency and robustness properties of predictors based on locally learnt support vector machines.

E0597: **The stability of SVMs**
Presenter: **Daohong Xiang**, Zhejiang Normal University, China
In the practical application of kernel based methods like SVMs, practitioner often chooses model parameters, such as regularization parameter and scaling parameter of the kernels, empirically. It is natural to ask what would happen on SVMs estimator when the data points deviate from the structure of the data set. We will present many kernel based methods like SVMs have nice stability properties if simultaneously the distribution, the regularization parameter and the kernel change slightly.

E0737: **Stability of pairwise learning methods**
Presenter: **Andreas Christmann**, University of Bayreuth, Germany
Pairwise kernel learning methods are often used to solve ranking problems and other non-standard problems beyond classification and regression. It will be shown that many pairwise learning methods based on kernels are not only statistically robust with respect to small changes of the probability measure or the data set, but that such methods have additionally nice stability properties with respect to small changes of the regularization parameter and with respect to the kernel or its reproducing kernel Hilbert space.

EO561 Room MAL 532 RECENT DEVELOPMENTS ON SPATIO-TEMPORAL MODELLING Chair: Sudipto Banerjee

E0378: **Bayesian model-based space-time joint interpolation of temperature and rainfall fields**
Presenter: **Giovanna Jona Lasinio**, Sapienza University of Rome, Italy
Co-authors: Gianluca Mastrantonio, Alessio Pollice, Carlo Blasi, Giulia Capotorti, Giulio Genova, Lorenzo Teodonio
A hierarchical space-time model is presented which describes the joint series of monthly extreme temperatures and amounts of rainfall. Data are available for 360 monitoring stations over 60 years with missing data affecting all series. Model components account for spatial dependence, seasonality, dependence on covariates (elevation) and between responses. The very large amount of data is tackled modelling spatial dependence by the nearest-neighbor Gaussian process. Response dependence is described introducing coregionalization. The proposed approach allows the imputation of missing data and an easy interpolation of climate surfaces at the national level. The motivation behind is the characterization of the so called ecoregions over the Italian territory. Ecoregions delineate broad and discrete ecologically homogeneous areas of similar potential as regards the climate, physiography, hydrography, vegetation and wildlife, and provide a geographic framework for interpreting ecological processes, disturbance regimes, and vegetation patterns and dynamics. In Italy two main macro-ecoregions have been defined and hierarchically arranged into 36 zones. Current climatic characterization of Italian ecoregions is based on data and bioclimatic indices from 1955-1985 and it requires appropriate update.

E1152: **Spatial disease mapping using directed acyclic graph autoregressive (DAGAR) model**
Presenter: **Abhirup Datta**, Johns Hopkins University, United States
Hierarchical models for regionally aggregated disease incidence data commonly involve region specific latent random effects which are modelled jointly as having a multivariate Gaussian distribution. The covariance or precision matrix incorporates the spatial dependence between the regions. Common choices for the precision matrix include the widely used intrinsic conditional autoregressive model which is singular, and its nonsingular extension which lacks interpretability. We propose a new parametric model for the precision matrix based on a directed acyclic graph representation of the spatial dependence. Our model guarantees positive definiteness and, hence, in addition to being a valid prior for regional spatially correlated random effects, can also be applied directly to model other dependent multivariate data like images and networks. Theoretical and empirical results demonstrate the interpretability of parameters in our model. Our precision matrix is sparse and the model is highly scalable for large datasets. We also derive a novel order-free version which remedies the dependence of directed acyclic graphs on the ordering of the regions by averaging over all possible orderings. The resulting precision matrix is still sparse and available in closed form. We demonstrate the superior performance of our models over competing models using simulation experiments and a public health application.

E1081: **Spatial modelling of fish counts in stream networks: Convolution or multilevel approaches**
Presenter: **Joao Pereira**, Universidade Federal do Rio de Janeiro, Brazil
Co-authors: Marco Rodriguez, Alexandra Schmidt, Julie Deschenes
In river networks, spatial correlation in species abundance arises from ecological processes operating at multiple spatial scales, including (1) directional aquatic processes constrained by waterflow, such as fish movement and transport of nutrients; (2) terrestrial processes that influence watershed characteristics at various spatial scales. A Poisson-lognormal mixture model of fish counts in a river network is considered. The mixing component accounts for temporal and spatial effects. Two different approaches for the spatial latent effects are explored: one based on a moving-average (convolution) construction, the other on a multilevel structure. The convolution approach can incorporate both hydrological ("as the fish swims") distance, to capture directional effects, and Euclidean ("as the crow flies") distance to capture effects of terrestrial processes. In contrast, the multilevel approach captures habitat effects nested at three discrete spatial scales (section, reach, and branch). Covariates were used to account

for local environmental effects. Bayesian inference is performed for a set of models representing different combinations of temporal and spatial effects. Models are compared by means of different information criteria, exam of the variance associated with each of the latent components, and prior-posterior comparison to clarify the relative contribution of different processes to variation in fish counts.

E1389: **Global estimation of air quality and the burden of disease associated with ambient air pollution**
Presenter: **Gavin Shaddick**, University of Exeter, United Kingdom

Ambient (outdoor) air pollution is a major risk factor for global health with an estimated 3 million annual deaths being attributed to fine particulate matter ambient pollution (PM2.5). The primary source of information for estimating exposures has been measurements from ground monitoring networks, however there are regions in which monitoring is limited. Ground monitoring data therefore needs to be supplemented with information from other sources, such as satellite retrievals of aerosol optical depth and chemical transport models. A hierarchical modelling approach for integrating data from multiple sources is proposed allowing spatially-varying relationships between ground measurements and other factors that estimate air quality. Set within a Bayesian framework, the Data Integration Model for Air Quality (DIMAQ) is used to estimate exposures, together with associated measures of uncertainty, on a high-resolution grid. Bayesian analysis on this scale can be computationally challenging and here Integrated Nested Laplace Approximations (INLA) are used. Estimated exposures from the model, produced on a high-resolution grid (10km x 10km) covering the entire globe and based on summaries of resulting the posterior distributions in each cell, it is estimated that 92% of the world's population reside in areas exceeding the World Health Organization's Air Quality Guidelines.

EO156 Room MAL 538 ADVANCES IN FUZZY CLUSTERING Chair: Maria Brigida Ferraro

E0492: **A fuzzy clustering based data fusion method**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

A method of data fusion based on fuzzy clustering is proposed. Recently, data fusion has gained tremendous interest as a means to combine different datasets in big data analysis. How to include the common variables though the different datasets is an important issue in data fusion. This study takes a new approach which utilizes fuzzy clusters obtained from fuzzy clustering through different data sets as the common variables in the data fusion. Then, the clusters for each dataset are used as a scale to select the variables from the combination of all variables of all the datasets to explain each original dataset. Also, the selected variables can be base spans of a linear subspace to explain the original each dataset. Since the linear subspace can be obtained for each original dataset, the multiple linear subspaces can be obtained for the multiple datasets. Then using a projector which projects data in different datasets to the intersection of these multiple linear subspaces, the single common predicted data values through different data sources can be obtained. This method will reduce big data to a space of lower dimensions, and obtains a unique solution through multiple datasets. Numerical examples will be included.

E0663: **Mixture-based clustering for the ordered stereotype model**
Presenter: **Daniel Fernandez**, Victoria University of Wellington, New Zealand
Co-authors: Richard Arnold, Shirley Pledger

Many of the methods that deal with clustering in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. Additionally, the use of ordinal data is very common (e.g. Likert or pain scale). Recent research has developed a set of likelihood-based finite mixture models for a data matrix of ordinal data. This approach applies fuzzy clustering via finite mixtures to the ordered stereotype model. Fuzzy allocation of rows, columns, and rows and columns simultaneously (biclustering) to corresponding clusters is obtained by performing a Reversible-Jump MCMC sampler. Examples with ordinal data sets will be shown to illustrate the application of this approach.

E1330: **Model-based and fuzzy clustering algorithms: A comparative simulation study**
Presenter: **Maria Brigida Ferraro**, Sapienza University of Rome, Italy
Co-authors: Marco Alfo, Paolo Giordani

Model-based and fuzzy clustering (algorithms) are widely used soft clustering methods. In both cases the objects are not assigned to the clusters via a hard rule allocation. In the former approach, it is assumed that the data are generated by a mixture of probability distributions (usually multivariate Gaussian) in which each component represents a different group or cluster. Each object is ex-post assigned to the clusters using the so-called posterior probability of component membership. In the latter case, no probabilistic assumptions are made and each object belongs to the clusters with a fuzzy membership degree, taking values in $[0,1]$, based on the distances between the objects and the cluster prototypes. Therefore it is quite obvious that the posterior probability of component membership may play a role similar to the membership degree. The aim is at comparing the performance of both approaches by means of a simulation study, also considering robust variants of both clustering approaches. In detail, in the model-based context, also finite mixtures of t distributions with or without trimming are investigated, whilst, in the fuzzy context, the problem of outliers is dealt by means of the so-called underlying noise cluster.

E0888: **Soft clustering of time series: New methods considering fuzzy, mixture models and probabilistic-D approaches**
Presenter: **Jose Vilar**, Universidade da Coruna, Spain
Co-authors: Borja Lafuente-Rego

New soft clustering algorithms to deal with time series considering fuzzy, nonparametric mixed models and probabilistic D-clustering techniques are introduced. In all cases, dissimilarity between series is measured in terms of squared Euclidean distance between sample quantile autocovariance vectors (QAD). The fuzzy model relies on a fuzzy C-medoids approach where QAD is used to calculate distances between series and medoids. Based on the asymptotic representation of the log-periodogram by means of a nonparametric regression model with log-exponentially distributed errors, an EM algorithm is used to estimate the components of a finite mixture model involving nonparametric approximations of the log-periodograms for each cluster. The probabilistic D-clustering technique is directly extended to the time series framework by considering QAD as a suitable feature-based distance between series. Supported by the nice properties of QAD, all the proposed algorithms show a very good clustering performance in simulations. Furthermore, the procedures are compared and the strengths and weaknesses of each one remarked and discussed.

EO569 Room MAL 539 COPULA-BASED REGRESSION Chair: Ingrid Hobaek Haff

E0548: **Extended D-vine quantile regression with applications**
Presenter: **Claudia Czado**, Technische Universitaet Muenchen, Germany
Co-authors: Daniel Kraus, Thomas Nagler

Non-Gaussian dependence patterns between the response and the covariates cannot be captured by ordinary quantile regression. In addition estimated quantile regression lines can cross for different levels, which is non desirable. Therefore we propose to use a tailored D-vine copula model to capture the dependence. It is tailored to select important covariates in a forward selection procedure by maximizing the conditional log likelihood. Extensions allow for non parametric pair copulas and discrete covariates in a parametric and non parametric setup. The approach will be introduced and illustrated in applications.

E0658: **Copula-based logistic regression estimation**
Presenter: **Taoufik Bouezmarni**, Universite de Sherbrooke, Canada

A new methodology is explored for estimating the success probability in a logistic regression setting. The idea consist in writing this later in terms of conditional copula and marginal distributions. The approach developing here consist, first, selecting a parametric family of copula that describes better the data at hand, and estimating non parametrically the marginal distributions, subsequently, we use the plug-in method to build an estimator the desired probability either in a binary case or categorical one. The availability of a rich family of copula, a various goodness-of-fit test and a nonparametric estimation of the marginal makes the approach more flexible. The asymptotic properties related to this estimators are provided. Finally, a simulated study is carried out to evaluate the performance of the newly proposed procedure, whether on simulated examples or real data.

E0595: **Gaussian copula regression**
Presenter: **Cristiano Varin**, Ca Foscari University of Venice, Italy
Co-authors: Guido Masarotto

A general framework for modelling dependence in regression models is presented based on a working Gaussian copula. Model properties, likelihood inference and computational aspects will be discussed in detail. Emphasis will be given to the validation of the Gaussian copula assumption and methods to make inference robust to departures from that assumption. The methodology will be illustrated with a variety of real data examples about time series, longitudinal and spatial data. Computations are based on the new version of the R package gcmr available through the CRAN archive.

E1077: **Multivariate frequency-severity regression models in insurance**
Presenter: **Lu Yang**, University of Amsterdam, Netherlands
Co-authors: Edward Frees, Gee Lee

In insurance and related industries including healthcare, it is common to have several outcome measures that the analyst wishes to understand using explanatory variables. For example, in automobile insurance, an accident may result in payments for damage to one's own vehicle, damage to another party's vehicle, or personal injury. It is also common to be interested in the frequency of accidents in addition to the severity of the claim amounts. We will synthesize and extend the literature on multivariate frequency-severity regression modeling with a focus on insurance industry applications. Regression models for understanding the distribution of each outcome continue to be developed yet there now exists a solid body of literature for the marginal outcomes. The focus is on the use of a copula for modeling the dependence among these outcomes; a major advantage of this tool is that it preserves the body of work established for marginal models. We illustrate this approach using data from the Wisconsin Local Government Property Insurance Fund. This fund offers insurance protection for (i) property; (ii) motor vehicle; and (iii) contractors' equipment claims. We find significant dependencies for these data; specifically, we find that dependencies among lines are stronger than the dependencies between the frequency and average severity within each line.

EO631 Room MAL 540 MISCELLANEOUS RESULTS FOR CHANGE POINT ANALYSIS Chair: Marie Huskova

E0296: **Detecting at-most-m changes in linear regression model**
Presenter: **William Pouliot**, University of Birmingham, United Kingdom
Co-authors: Lajos Horvath, Shixuan Wang

A new procedure is provided to test for at most two changes in the time dependent regression model $y_t = x_t\boldsymbol{\beta}_t + e_t$ with $1 \leq t \leq T$. Our procedure is based on weighted sums of the residuals, incorporating the possibility of two changes. The weak limit of the proposed test statistics the sum of two double exponential random variables. A small Monte Carlo simulation illustrates the applicability of the limit results in case of small and moderate sample sizes. We compare the new method to the CUSUM and standardized (weighted) CUSUM procedures and obtain the power curves of the test statistics under the alternative. We apply the method to find changes in the unconditional four factor CAPM.

E0913: **Nonparametric changes in variance detection using localised estimates**
Presenter: **Rebecca Killick**, Lancaster University, United Kingdom
Co-authors: Jamie-Leigh Chapman, Idris Eckley

A nonparametric method of detecting changes in variance is developed for the case where assumptions of normality and independence are not appropriate. This may be the case in applications in renewable energy, finance and biomedical research. We develop a nonparametric method using the Locally Stationary Wavelet (LSW) model to provide a local estimate of the variance of a time series. If a time series has constant variance, then its local variance function will have constant mean and variance over time. However, if this variance is piecewise constant, then the structure of the variance function will also be piecewise constant. To this end, we can use the wavelet transformation to identify changes in variance in a fully non-parametric setting. We will demonstrate the efficacy of our approach through simulations studies where we compare against the most commonly used approaches in a variety of settings. The proposed method performs particular well in cases where the data contains outliers, heavy tails and/or autocorrelation. As the wavelet transformation separates out the variance from the autocorrelation the methodology is very fast compared to other approaches where the ARMA likelihood is required to account for the autocorrelation in a series. We demonstrate this on a large real world example.

E1329: **Change detection and inference in high-dimensional covariance matrices based on l_1- and l_2-projections**
Presenter: **Ansgar Steland**, University Aachen, Germany
Co-authors: Rainer von Sachs

New results about inference and change point analysis of high dimensional vector time series are discussed. The results deal with change-point procedures that can be based on an increasing number of bilinear forms of the sample variance-covariance matrix as arising, for instance, when studying change-in-variance problems for projection statistics. Areas of applications are shrinkage, aggregated sensor data and dictionary learning. Contrary to many known results, e.g. from random matrix theory, the results hold true without any constraint on the dimension, the sample size or their ratio, provided the weighting vectors are uniformly l_1-bounded. Extensions to a large class of l_2-bounded projections are also discussed. The large sample approximations are in terms of (strong resp. weak) approximations by Gaussian processes for partial sum and CUSUM type processes. It turns out that the approximations by Gaussian processes hold not only without any constraint on the dimension, the sample size or their ratios, but even without any such constraint with respect to the number of bilinear form. For the unknown variances and covariances of these bilinear forms nonparametric estimators are proposed and shown to be uniformly consistent. We also discuss how the theoretical results lead to novel distributional approximations and sequential methods for shrinkage covariance matrix estimators.

E0584: **Some results on detection of change points in panel data**
Presenter: **Marie Huskova**, Charles University, Czech Republic
Co-authors: Jaromir Antoch, Jan Hanousek, Zdenek Hlavka, Lajos Horvath, Shixuan Wang

Detection of changes in the panel data setup is considered when T (time dimension) is finite/moderate and N (number of panels) tends to infinity. Some test procedures (no change versus there is a change) and estimators of the time of a change. A proper wild bootstrap procedure is proposed. The asymptotic behavior of the proposed test statistics and estimators is investigated. Theoretical results are accompanied by simulations and the

application testing procedure in the framework a CAPM model.

EO447 Room MAL 541 TIME SERIES AND REGRESSION MODELS Chair: Marco Meyer

E1237: **Estimating non-causal VARs using multivariate all-pass filters**
Presenter: **Bernd Funovits**, University of Helsinki, Finland

An estimation strategy is proposed based on multivariate all-pass filters for a possibly non-causal VAR system driven by non-Gaussian i.i.d. shocks. Multivariate all-pass filters are matrices whose elements are rational functions and which preserve the second moment properties of the solution of the (transformed) VAR system. Applying a (non-trivial) all-pass filter to a non-Gaussian i.i.d. process results in an uncorrelated (white noise) process which is, however, not independent across time. First, we obtain the finite set of VAR systems whose solutions have the same second moment properties but whose determinantal roots are different. Subsequently, we minimize an objective function (using higher moments) which is minimal for the true VAR system.The obtained estimator can serve as an initial estimate for maximum likelihood estimation.The algorithm is implemented in the R-package varAllPass.

E1232: **Adaptive bandwidth selection for M-estimators in locally stationary time series**
Presenter: **Stefan Richter**, Heidelberg University, Germany

A general class of locally stationary processes which are assumed to depend on an unknown finite-dimensional parameter curve is studied. For the estimation of these curves, we use nonparametric local M-estimators depending on a bandwidth. We investigate a local bandwidth selection procedure via contrast minimization. We prove that the corresponding estimators for the parameter curves achieve the asymptotically optimal minimax rate up to a log factor. All important conditions of our theorems only concern the (usually well-known) corresponding stationary time series model which allows for an easy verification. The performance of the method is studied in a simulation for some examples like tvAR, tvARCH and tvMA processes.

E0963: **Linear regression with time series regressors**
Presenter: **Suhasini Subbarao**, Texas A&M, United States

In several diverse applications, from neuroscience to econometrics, it is of interest to model the influence regressors have on a response. In many of these applications, the regressors have a meaningful ordering (usually a time series), but the number of regressors is very large. Linear regression, where the number of regressors n is of the same order or magnitudes larger than the observed number of responses p has received considerable attention in recent years. However, most of these approaches place a sparsity assumption on the regressor coefficients. When the regressors are a time series, the sparse assumption can be unrealistic with no intuitive interpretation. We consider the problem of linear regression with stationary time series regressors, but work under the weaker assumption that the regressor coefficients are absolutely summable. We propose a computationally efficient method for consistently estimating the regression parameters, that avoids matrix inversion.

E1280: **Extending the range of validity of autoregressive bootstrap methods**
Presenter: **Efstathios Paparoditis**, University of Cyprus, Cyprus

Two modifications of the autoregressive-sieve respectively autoregressive bootstrap are proposed. The first modification replaces the classical i.i.d. resampling scheme applied to the residuals of the autoregressive fit by a generation of i.i.d. wild pseudo-innovations that appropriately mimic the first, the second and the fourth order moment structure of the true innovations driving the underlying linear process. This modification extends the validity of the autoregressive-sieve bootstrap to classes of statistics for which the classical, residual-based autoregressive-sieve bootstrap fails. In the second modification, an autoregressive bootstrap applied to an appropriately transformed time series is proposed which, together with a dependent-wild type generation of pseudo-innovations, delivers a bootstrap procedure which is valid for large classes of statistics and for stochastic processes that satisfy quite general weak dependent conditions. A fully data-driven selection of the bootstrap parameters involved in both modifications is proposed, while extensive simulations, including comparisons with alternative bootstrap methods, show a good finite sample performance of the proposed bootstrap procedures.

EC699 Room MAL 152 CONTRIBUTIONS IN HYPOTHESIS TESTING Chair: Clement Marteau

E1048: **A new class of tests for multinormality based on the moment generating function**
Presenter: **Jaco Visagie**, University of Pretoria, South Africa, South Africa
Co-authors: Norbert Henze

It is often of interest to determine whether or not observed data are realised from a multivariate normal distribution. We propose a new class of affine invariant tests for multinormality based on the empirical moment generating function. The proposed test statistic contains a tuning parameter. We show that, after suitable scaling, the test statistic approaches a linear combination of two well-known measures of multivariate skewness for fixed sample sizes as the tuning parameter tends to infinity. We also derive the limit distribution of the test statistic for large samples under the null hypothesis of multinormality. A Monte Carlo power study is conducted in order to compare the finite sample performance of the proposed test to various existing tests.

E1690: **Multivariate mean testing under block exchangeable covariance structure**
Presenter: **Ivan Zezula**, P.J. Safarik University, Slovakia
Co-authors: Daniel Klein, Anuradha Roy

A multivariate linear normal model with block exchangeable covariance structure is considered. Such a structure can be naturally implied by the design of the experiment. Its main advantage is strongly reduced number of the second order parameters, which allows mean testing even with a small sample size. We will compare performance of several test statistics used in this setting. Methods will be demonstrated on real dataset.

E0292: **Testing the equality of coefficients of variations with right-censored data**
Presenter: **Nihan Potas**, Gazi University, Turkey

The coefficient of variation is commonly used as a measure of relative variability in clinical trials and medical sciences. To combine or compare variability from several clinical trials has become necessary. Under these circumstances, parametric or non-parametric tests for the equality of coefficient of variations must be widely considered. In the past most of the papers concentrated upon testing coefficient of variations from k normal distributed populations. However, for most of the clinical trials, medical sciences data are longitudinal, failure time observation and it is unlikely to omit the censoring cases. According to that, the aim is to present the test statistics for right censored data to test the equality of coefficient of variations. The empirical sizes and powers of the tests statistics are computed and compared with a simulation study. We provide examples using right censored survival data from oncology clinic. Based on the simulation results right censoring case, likelihood ratio tests give quite good results compare to score tests.

E1680: **Statistical inference for random coefficient dynamic panel data models**
Presenter: **Remigijus Leipus**, Vilnius University, Lithuania
Co-authors: Anne Philippe, Donatas Surgailis, Vytaute Pilipauskaite
The purpose is to discuss nonparametric estimation of the distribution function of the autoregressive coefficient from a panel of random coefficient AR(1) data. Consistency and asymptotic normality of the empirical distribution function and a class of kernel density estimators, as number and length of panels increase to infinity, is established. The Kolmogorov-Smirnov goodness-of-fit test for simple and composite hypotheses of Beta distributed coefficient is discussed. A simulation study for goodness-of-fit testing examines the finite-sample performance. The corresponding tail-index estimator and its properties are provided.

EG026 Room MAL 151 CONTRIBUTIONS TO TIME SERIES ANALYSIS I Chair: Rajendra Bhansali

E1426: **A bootstrap contribution in STL decomposition**
Presenter: **Clara Cordeiro**, CEAUL and FCT, University of Algarve, Portugal
Co-authors: Manuela Neves, M Rosario Ramos
A time series is the result of observing the values of a variable over time during regular intervals and can be seen as the result of the combination of different components. The classical methods of decomposition of time series allow to identify the trend, the seasonality and the irregular components. However, these methodologies do not allow for a flexible specification of the seasonal component, and the trend component is generally represented by a deterministic time function, which is easily affected by the existence of outliers. The nonparametric Seasonal-Trend decomposition by Loess (STL) is able to identify a seasonal component that changes over time, a non-linear trend and it can be robust in the presence of outliers. Bootstrap methods, initially introduced for independent random variables, can be successfully applied to time series. The Boot.EXPOS procedure, joining bootstrap and exponential smoothing methods, has revealed promising results. The aim is to explore the use of the Boot.EXPOS in predicting the components of the STL decomposition. Based on an error measure, the best STL fit is chosen. In the case of an uncorrelated irregular component, the forecast of the STL will rely on the forecast of trend and seasonal components obtained through the Boot.EXPOS.

E1436: **Estimation of autoregressive models from data with measurement error**
Presenter: **Jessa Lopez**, University of the Philippines-Diliman, Philippines
When used in modeling, measurement error in the data influences the model structure, and it can even distort the entire data generating process. A method based on spline smoothing and bootstrap is proposed to estimate an autoregressive process to address the issues associated with data contaminated with measurement error. The simulation study indicates that the method estimates the parameters of an autoregressive model better than the usual conditional least squares estimate when the data is measured with error.

E1739: **Nonparametric Gaussian inference for stable processes**
Presenter: **Fabian Mies**, RWTH Aachen University, Germany
Co-authors: Ansgar Steland
Jump processes driven by α-stable Levy processes impose inferential difficulties as their increments are heavy-tailed and the intensity of jumps is infinite. We consider the estimation of the functional drift and diffusion coefficients from high-frequency observations of a stochastic differential equation. By transforming the increments suitably prior to a regression, the variance of the emerging quantities may be bounded while allowing for identification of drift and diffusion in the limit. The findings are applied to obtain a comprehensive treatment of the asymptotics of a nonparametric kernel estimator, covering asymptotic normality and consistency of subsampling approximations, and of a parametric volatility estimator for the Ornstein-Uhlenbeck process. The proposed approach also suggests a semiparametric estimator for the index of stability α.

E1151: **Optimal estimation of change-points in time series**
Presenter: **Chun Yip Yau**, Chinese University of Hong Kong, Hong Kong
Asymptotic theory is established for the optimal estimation of change-points in time series. We show that the Bayes estimator is asymptotically efficient for change-point estimation under mean squared error loss. Two subsampling procedures are developed to construct confidence intervals for the change-points. Simulations and real data applications are presented to investigate the finite sample performance of the Bayes estimator and the two subsampling procedures.

CO504 Room Bloomsbury NONPARAMETRIC METHODS IN MICROECONOMETRICS Chair: Christoph Breunig

C0675: **Simultaneous confidence bands for ratios of quantile functions and growth incidence curves**
Presenter: **Fabian Dunker**, University of Canterbury, New Zealand
Co-authors: Tatyana Krivobokova, Stephan Klasen
Ratios of quantile functions are an important tool to evaluate the distributional pattern of growth processes when repeated cross-sectional data are available. The most popular example are Growth Incidence Curves (GIC) that allow assessments whether income growth in developing countries has been pro-poor. Such assessments depend on the location and slope of the growth incidence curve as well as the confidence bands surrounding the curve. We present a construction of uniform confidence bands for GICs and similar ratios of quantile functions. In contrast to existing point-wise confidence bands that are created via bootstrapping, the bands we propose are valid simultaneously for almost all points in the domain of the quantile functions and GICs. They allow for an assessment of the location and on larger scales of the slope of the curves. Furthermore, the construction does not depend on bootstrapping but on an analysis of the asymptotic distribution of GICs. This allows for significantly faster algorithms. The performance of the confidence band is demonstrated in simulations and in an example using income data from Uganda for 1999-2005.

C0982: **Confidence sets for group memberships**
Presenter: **Andreas Dzemski**, University of Gothenburg, Sweden
Co-authors: Ryo Okui
Panel models are considered where unit behavior is driven by a latent group structure. New procedures are proposed for constructing confidence sets for group memberships. We consider unitwise confidence sets as well as confidence sets for the entire group membership structure, which we call uniform confidence sets. A unitwise confidence set gives a set of possible group memberships for one specific unit and contains that units true group membership with a pre-specified probability. A uniform confidence set gives a set of possible group memberships for each unit that contains the true group memberships with a pre-specified probability. It is constructed by inverting a test that tests group memberships for all units simultaneously. Our confidence sets can be used to quantify the uncertainty about estimated group memberships. This complements previous work that focuses on inference with respect to the parameters that govern group-specific behavior. Our approach exploits the fact that the problem of sorting units into groups can be characterized by a system of moment inequalities. We construct the uniform confidence sets by stringing together unitwise confidence sets using a Bonferroni correction. The theoretical justification of this procedure exploits a high-dimensional CLT and a new anti-concentration result. We also propose an algorithm that combines moment selection with iterated hypothesis selection to eliminate units for which group membership can be precisely estimated.

C0682: **Heterogeneous employment effects of job search programmes: A machine learning approach**
Presenter: **Michael Lechner**, University St Gallen, Switzerland
Co-authors: Michael Knaus, Anthony Strittmatter
The effect heterogeneity of job search programmes for unemployed workers is systematically investigated. For that, we combine non-experimental causal empirical models with Lasso-type estimators. The empirical analysis is based on rich administrative data from Swiss social security records. We find considerable heterogeneities only during the first six months after the start of training. Consistent with previous results of the literature, unemployed with lower employment opportunities profit more from participating in these programmes. Furthermore, we also document heterogeneous employment effects by immigration status. Finally, we show the potential of easy-to-implement programme participation rules for improving average employment effects of these active labour market programmes.

C1176: **Estimation and inference of treatment effects with L_2-boosting in high-dimensional settings**
Presenter: **Martin Spindler**, University of Hamburg, Germany
Boosting algorithms are very popular in Machine Learning and have proven very useful for prediction and variable selection. Nevertheless in many applications the researcher is interested in inference on treatment effects or policy variables in a high-dimensional setting. Empirical researchers are more and more faced with rich data sets containing very many controls or instrumental variables where variable selection is challenging. We give results for valid inference of a treatment effect after selecting amongst very many control variables and instrumental variables estimation with potentially very many instruments when post- or orthogonal L_2-Boosting is used for variable selection. We give simulation results for the proposed methods and an empirical application.

CO512 Room Chancellor's Hall TIME SERIES ECONOMETRICS Chair: Antonio Montanes

C0344: **Polar warming**
Presenter: **Lola Gadea**, University of Valencia, Spain
Co-authors: Jesus Gonzalo Munoz
It is well known that what happens in the Arctic does not stay in the Arctic; but we need to know what is really happening there. We focus on the temperature in the Arctic circle. Applying a previous methodology, Polar Warming is defined as an increasing trend in certain distributional characteristics (moments, quantiles, etc) of polar temperatures, and not only on the average. We consider Arctic temperatures as a functional stochastic process from which we obtain distributional characteristics as time series objects. We apply our robust trend test to three different temperature databases that contain data from different stations in the Arctic: Climate Research Unit (CRU), Arctic Air Surface Temperature (SAT) and ERA-Interim reanalysis form the European Centre for Medium-Range Weather Forecast (ECMWF). The first two run from the ninetieth century to nowadays and the last one from 1979-2017. Preliminary analyses show that, with the first two, we obtain similar results to the ones obtained when analyzing the global temperature: (i) there is a clear positive trend in all the distributional characteristics of the temperature variable; (ii) the lower quantiles have a larger trend than the mean and upper quantiles and (iii) the variability is decreasing (variance has a negative trend, lower temperatures approach the median faster than the higher ones deviate from it).

C0339: **External imbalances and economic growth across countries: A non-parametric reassessment**
Presenter: **Mariam Camarero**, University Jaume I, Spain
Co-authors: Cecilio Tamarit, Jesus Peiro-Palomino
The role played by imbalanced net foreign asset (NFA) positions on growth is assessed. This topic is in the spotlight in recent years in both the theoretical and the empirical spheres. The literature, however, is still inconclusive, inviting authors to depart from one-size-fits-all prescriptions and to adopt flexible approaches allowing for parameter heterogeneity. We revisit the external imbalances-growth nexus by running non-parametric kernel regressions for a wide sample of countries and for the period 1983-2011, controlling also for well-established growth determinants and providing results robust to endogeneity. The main results suggest a preponderant positive relationship between NFA and growth, although the impact varies across countries and over time. In particular, it is affected by country-specific factors such as the depth of financial development and the quality of formal institutions. However, the intensity of the causal relationship is not statistically different for debtor and creditor countries, or for fast and slow growing countries.

C1266: **Cointegration, cobreaking and cotrending for trending time series**
Presenter: **Josep Lluis Carrion-i-Silvestre**, Universitat de Barcelona, Spain
A set of variables with two types of non-stationary features, stochastic trends and broken linear trends is considered. We develop tests that can be used to see if there exists a linear combination of these variables under which the non-stationary features can be cancelled out. We provide a comprehensive treatment of cointegration testing with structural breaks that is missing in the literature. In particular, the first test we develop can be used to see if stochastic trends can be eliminated and thus cointegration holds, regardless of whether structural breaks in linear trends are eliminated. The second test can be used to see if both stochastic trends and breaks in linear trends are simultaneously removed and thus both cointegration and cobreaking simultaneously hold. The third test can be used to see if not only breaks in linear trends but also linear trends themselves are eliminated along with stochastic trends and thus both cointegration and cotrending hold. We provide the asymptotic null distributions of proposed test statistics and some Monte Carlo simulation results to assess the adequateness of our asymptotic distributions in samples with common sizes.

C1325: **Testing for cointegration under the presence broken trends**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain
Co-authors: Lola Gadea
The asymptotic behavior of a previous statistic when the variables being studied may exhibit broken trends is studied. We prove that this statistic works very poorly in these circumstances, leading to wrong results when the variables are I(1) or I(0) plus broken trends. We propose a method in order to correct this problem, based on the appropriate extraction of the deterministic elements. Our Monte Carlo results are promising, showing that our method works properly in finite samples.

CO593 Room Court DURHAM QUANT-FIN GROUP SESSION: HIGH FREQUENCY TRADING AND ECONOMETRICS Chair: Julian Williams

C0392: **Spanning GARCH: Pricing uncertainty in the very long run**
Presenter: **Abderrahim Taamouti**, Durham University Business School, United Kingdom
High-order moments derivatives, such as variance and skewness swaps, are increasingly popular risk management tools for Foreign Exchange exposures. We propose a simple Fourier inversion-based method to estimate the risk neutral value of a spanning contract contingent on the future outcome of an asset price. The price of the contract can computed by a single quadrature evaluation of the characteristic function and is easily implemented for moment swaps beyond variance, such as skewness and kurtosis. Indeed, even fractional moment swaps are admissible in this framework and can be estimated directly from spot and yield curve data via maximum likelihood. We show that the predicted moment swaps based on our approach - with the characteristic function computed using a GARCH model- perform well against both option implied variance and skewness swaps and traded variance swaps.

C1093: **Credit, stock, commodity and shipping market interactions in a three state boom bust cycle**
Presenter: **Nikos Paltalidis**, Durham University, United Kingdom

A three-regime boom-bust cycle model is proposed which captures short-term transmission of shocks to the shipping market. Stock, credit and commodity assets are employed as proxies that capture short-run market fluctuations for dry bulk, and tanker shipping markets. The results indicate that during the latest boom-bust cycle from 2005 till 2013, there is evidence of cross-border volatility spillover among seemingly unrelated assets which behave as channels that spread risk, and act as an indicator for the shipping cycle phase. Higher risk premium required by investors for holding financial assets, depresses substantially and changes the risk perception of the shipping market. We document that these assets become more integrated during crisis periods when compared with tranquil periods. However, in the post-crisis era supply side factors dictate the fluctuations of the shipping assets.

C1030: **Recovering foreign exchange option volatility surface from spot price dynamics**
Presenter: **Handing Sun**, Durham University, United Kingdom
Co-authors: Julian Williams, Dennis Philip, Julian Cook

Over-the-counter foreign exchange options (OTC-FXOs) are the fourth largest derivatives market in the world. However, the extant literature on their pricing is noticeably thin. We propose a new discrete time exponential- affine multi-factor model, with multiple estimation strategies and pricing confidence intervals for the resulting synthetic volatility surface, for OTC- FXOs and test the various specifications out-of-sample on five liquid currencies versus the dollar. Our specification is the first of its type that can be estimated directly from spot FX and deposit rate quotes without recourse to quoted volatility surfaces. Results indicate that both short and long tenor OTC-FXOs can be accurately priced with minimal calibration.

C1031: **The information content of option implied moments and co-moments**
Presenter: **Julian Williams**, Durham University, United Kingdom
Co-authors: Yang Zhang

The aim is to introduce a new tensor derived computation of higher dimensional average implied correlations determined from high frequency options panels. Using supersymmetric identities to decrease we can determine average moments and co-moments for equities and incorporate these forward looking measures to a standard asset pricing framework. To compute the metrics we use both a broad market index (S&P 500) and nine sector indices versus the underlying components. Our empirical analysis shows that the cross section stock returns have substantial exposure to risk captured by the market average correlation factor. The results are robust to various permutations of the estimation procedure. The risk premium of the market average correlation factor is statistically and economically significant when controlling the other common market risk factors or firm characteristics. Finally, we test the higher-order CAPM with the ex ante market beta, gamma, and theta approximated by the option- implied average correlations. In line with the evidence documented in previous literature, we find positive significant risk premiums for the ex ante market beta and theta but mixed results for the ex ante market gamma.

CO638 Room G11 FINANCIAL ECONOMETRICS Chair: Roderick McCrorie

C1277: **The fractionally cointegrated VAR model with threshold adjustment**
Presenter: **Chi Wan Cheang**, University of Southampton, United Kingdom

The fractional cointegrated vector autoregressive (FCVAR) model is extended by allowing two regimes in the speed of adjustment parameter in the error-correction term, treating the long-run cointegrating vector, the fractional and cofractional orders invariant across regimes. Since the threshold parameter is not identified under the null hypothesis of no threshold, a SupLM test for the presence of a threshold is proposed for the FCVAR model. Hence the bootstrap null distribution and p-value are derived. The size and power of the test are investigated through a Monte Carlo simulation. The asymptotic of the cointegrating vector and the threshold parameter is discussed through the smooth transition case. The proposed nonlinear FCVAR model is applied on the volatility index (VIX) and its related futures. By the fact that investors treat the VIX-related products as a security on tailed risk, the VIX products are more attractive in uncertain time that in quiet time. It is argued that the adjustment between volatility index and its futures towards equilibrium could be regime-switching. Empirical result provides a theoretical ground to explain the momentum strategy used by volatility traders.

C1201: **Bubble migration across asset classes during the global financial crises**
Presenter: **Isabel Figuerola-Ferretti**, ICADE, Spain
Co-authors: Roderick McCrorie, Ramon Bermejo Climent, Gonzalo Suarez

A new, mildly explosive/multiple bubbles technology is combined with a bubble migration test to analyse the time series behaviour of a number of key macroeconomic and financial variables during the Global Financial Crisis (GFC). Following a recent approach, we use a dry cargo single ocean freight rate in our case the Baltic Dry Index (BDI) as a gauge of the economy. An anticipatory measure, it captures near future demand shifts for industrial commodities in global markets. In the financial sector, we document relevant mildly explosive episodes/bubbles in credit default swap (CDS) portfolios, high yield bond portfolios, the ICE Libor USD 6 month rate and other commodity prices. We shall also detect such an episode in the real economy, where U.S. jobless claims showed a statistically significant, mildly explosive rise before the Lehman Brothers episode that many feel delineated the highpoint of the GFC. Our work provides a richer description of the macro and financial aspects of the GFC and brings to it other dimensions than have been offered hitherto.

C1205: **A market approach for convergence trades**
Presenter: **Tao Tang**, Jinan University, China
Co-authors: Isabel Figuerola-Ferretti, Ioannis Paraskevopoulos

A VECM representation is proposed for cointegrated assets in the continuous-time framework that embeds the setup recently proposed in the literature. This model implies a simple framework to check for cointegration based on the condition that positive convergent speed ensures the stationarity of price spreads. A pair of cointegrated assets is then identified to derive a dynamically optimal pairs trading portfolio with a risk-free bond. This involves maximizing the portfolio value at terminal time without the requirement of a functional form for investors preferences. To do so, we connect the derived optimal portfolio with European-type spread options and in consequence the optimal investment policies can be modeled using the spread option's resulting delta hedging strategies. Our framework is tested empirically using pairs identified from the Dow Jones Industrial Average. We find that the proposed optimal strategy achieves a significant improvement in the Sharpe ratios and abnormal returns with respect to the previous work in the literature.

CO660 Room G3 THE ECONOMETRICS OF MARKET EXUBERANCE Chair: Florian Ielpo

C0229: **An empirical approach to financial crisis indicators based on random matrices**
Presenter: **Antoine Kornprobst**, University Paris Sorbonne, France
Co-authors: Raphael Douady
The aim is to build financial crisis indicators based on spectral properties of the dynamics of market data. After choosing an optimal size for a rolling window, the historical market data in this window is seen every trading day as a random matrix from which a covariance and a correlation matrix are obtained. The financial crisis indicators that we have built deal with the spectral properties of these covariance and correlation matrices and they are of two kinds. The first one is based on the Hellinger distance, computed between the distribution of the eigenvalues of the empirical covariance matrix and the distribution of the eigenvalues of a reference covariance matrix representing either a calm or agitated market. The idea behind this first type of indicators is that when the empirical distribution of the spectrum of the covariance matrix is deviating from the reference in the sense of Hellinger, then a crisis may be forthcoming. The second type of indicators is based on the study of the spectral radius and the trace of the covariance and correlation matrices as a mean to directly study the volatility and correlations inside the market. The idea behind the second type of indicators is the fact that large eigenvalues are a sign of dynamic instability. The predictive power of the financial crisis indicators in this framework is then demonstrated by using them as decision-making tools in a protective put strategy to hedge long ETF positions.

C1809: **Detecting end-of-sample explosive behaviour using persistence change tests**
Presenter: **Sam Astill**, University of Essex, United Kingdom
The purpose is to explore the possibility of utilising persistence change tests to detect a change in persistence in an economic time series from being an I(1) process to an explosive process at the end of the sample. The first differences of a series that is subject to a change in persistence from I(1) to explosive behaviour at the end of the sample is itself a series that follows an I(0) process with a shift to explosive behaviour at the end of the sample. As such, we examine the ability of tests for a change in persistence of a series from being I(0) to I(1) to detect explosive end-of-sample bubbles when applied to the first differences of a financial time series. We find, unsurprisingly, that persistence change tests reject for explosive alternatives in addition to the I(1) alternative for which they are designed. We also find that persistence change tests have good power to detect short lived end of sample explosive episodes as well as ongoing episodes of a longer duration. A test with such power properties will be desirable for practitioners, as existing tests are usually best suited to detecting explosive episodes of a particular duration.

C1457: **On unit root methods to predict financial crises**
Presenter: **Eero Tolo**, Bank of Finland, Finland
Co-authors: Timo Virtanen, Matti Viren, Katja Taipalus
Cross-country evidence is provided for the performance of unit-root-based early warning systems in ex-ante prediction of financial crises in 15 EU countries over the past three decades. We find, especially high performance for time series, that are explicitly related to debt, which issue signals a few years in advance of a crisis. Combining signals from multiple time series further improves the predictions. Our results suggest an early warning tool based on unit root methods provides a convenient accessory for financial stability supervision.

C1579: **Fundamental bubbles in equity markets**
Presenter: **Florian Ielpo**, Centre Economie de la Sorbonne, France
Co-authors: Mikita Kniahin
Using an affine model to compute the price of equities based on a dataset of macroeconomic factors, we propose a measure of "fundamental bubbles". The joint dynamics of macroeconomic factors and short rates in a VAR model allows for a mutual feedback mechanism between monetary policy and macroeconomic variables. We use a dynamic affine term structure framework to price equity and bonds, and investigate how prices are related to macro variables that span them. We analyse the discrepancies between market and model implied equity prices, the latter approximating "fundamental valuation". By testing the stationarity of the discrepancies, we investigate whether equity prices exhibit a bubble behavior. We perform the analysis over 3 major US and 3 major European equity indices, diagnose bubbles and confirm and its consequences for S&P500 and Dow Jones over the 1990-2017 period.

CO514 Room G4 TIME SERIES MODELS OF COMMODITIES AND COMMODITIES FUTURES Chair: Pierre Siklos

C0861: **Economic impact of commodity financialization**
Presenter: **Devraj Basu**, Strathclyde Business School, United Kingdom
The impact of financialization on commodity futures return distributions is studied. Over the period associated with financialization, we find shifts in the nature of volatility across the entire cross-section of commodities and shifts in the nature of the entire returns distribution across the agricultural and metals sectors. The effects of financialization appear to be the most profound on the metal sector and the altered nature of the metals return distributions appears to have transmitted itself to the mining sector, leading to simultaneous increases in capital expenditures which may have contributed to the recent value destruction in this sector.

C1005: **The quantile-heterogeneous autoregressive model of realized volatility: New evidence from commodity markets**
Presenter: **Konstantin Kuck**, University of Hohenheim, Germany
Co-authors: Robert Maderitsch
A cross-asset perspective is provided on state-dependence in the dynamics of realized volatility in the commodity futures market. Using high-frequency data for futures on Gold, Silver and Light Sweet Crude Oil, covering the period from 2007 to 2016, we estimate various Quantile-Heterogeneous Autoregressive models of daily realized volatility (Q-HAR-RV). The daily volatility is modeled as a linear function of own lags measured over different time resolutions to specifically account for the heterogeneous impact of market participants with different trading motives and investment horizons. Furthermore, using quantile regression, we are able to identify potential state-dependence and asymmetry in the short-, mid- and long-term autoregressive dynamics with respect to different volatility levels. Overall, the daily and monthly volatility aggregates appear more important compared to weekly volatility. However, we also document considerable changes in the relative importance of mid- and long-term volatility components under varying market conditions which appear remarkably similar across the three assets. Specifically, the impact of the weekly volatility increases distinctly from lower to higher quantiles of the conditional volatility distribution while that of daily and monthly volatility decreases. This implies that information generated over the medium-term gains importance in phases of increased uncertainty.

C0826: **Speculation and volatility: A time-varying approach applied on Chinese commodity futures markets**
Presenter: **Claudia Wellenreuther**, University of Muenster, Germany
Co-authors: Jan Voelzke
There is an ongoing discussion and a lot of empirical investigations on, whether or not, speculative activity increases volatility on commodity futures markets. However, relatively little empirical research has been conducted to analyze the role of speculators on Chinese commodity futures markets, mostly due to a missing standard measure for speculative activity. Additionally, most of the existing studies assume the potential effects to be constant over a long period of time. We address both shortcomings and empirically investigate the time-varying influence of speculative

activity on returns volatility in Chinese futures markets for commodities. To measure speculative activity a speculation ratio, defined as trading volume divided by open interest, is used. We sequentially apply two time-varying VAR models with stochastic volatility to six heavily traded metal and agricultural futures contracts to show how the relationship between returns volatility and the speculation ratio evolves over time. We analyze Granger causality as well as impulse responses. Eventually, we find no evidence for an effect from speculative activity on volatility. On the contrary, for most commodities, return volatility seems to amplify speculation.

C1417: **On crude oil as financial asset: Evidence from ten years of financialization**
Presenter: **Maria Kartsakli**, University of St. Gallen - Swiss Institute of Banking and Finance, Switzerland
The financialization of crude oil markets over the last decade has changed the behavior of oil prices in fundamental ways. We uncover the gradual transformation of crude oil from a physical to a financial asset. Although economic demand and supply factors continue to play an important role, recent indicators associated with financialization have emerged since 2008. We show that financial variables have become the main driving factors explaining the variation in crude oil returns and volatility today. Our findings have important implications for portfolio analysis and for the effectiveness of hedging in crude oil markets.

CO252 Room G5 EMPIRICAL MACRO Chair: Michael Owyang

C0197: **Trade flows, trade policy, and the size of fiscal multipliers**
Presenter: **Nora Traum**, North Carolina State University, United States
Co-authors: Matteo Cacciatore
The aim is to study how trade linkages affect the domestic and international transmission of fiscal shocks. We develop a two-country model with heterogeneous firms, endogenous producer entry in domestic and export markets, and a richly specified fiscal sector. We estimate the model with Bayesian methods and U.S. and Euro Area data. The model offers a fresh vantage on the international transmission of fiscal shocks. Fluctuations in the number of traded varieties embedded in trade flows quantitatively affect the size of government spending multipliers and spillover effects abroad. We show the strength of this channel depends crucially on expectations of how government debt is financed. We use the estimated model to provide a quantitative assessment of the effectiveness of a joint fiscal and trade policy intervention on output, trade, and government debt.

C0268: **Debt and stabilization policy: Evidence from a Euro area FAVAR**
Presenter: **Laura Jackson Young**, Bentley University, United States
Co-authors: Michael Owyang, Sarah Zubairy
The Euro-area poses a unique problem in evaluating counter-cyclical policy: A currency union with a shared monetary policy and country-specific fiscal policy. Analysis can be further complicated if high levels of public debt affect the performance of stabilization policy. We construct a framework capable of handling these issues with an application to Euro-area data. In order to incorporate multiple macroeconomic series from each country but, simultaneously, treat country-specific fiscal policy, we develop a hierarchical factor-augmented VAR with zero restrictions on the loadings that yield country-level factors. Monetary policy, then, respond to area-wide conditions but fiscal policy responds only to country-level conditions. We find that there is broad variation in different countries' responses to area-wide monetary policy and country-specific fiscal policy. Moreover, we find that debt conditions do not diminish the effectiveness of policy in a significant manner, suggesting that any negative effects must come through other channels.

C0295: **Gender differences in business cycle dynamics and policy implications**
Presenter: **Amy Guisinger**, Lafayette College, United States
Co-authors: Tara Sinclair
While traditional macroeconomic models take individuals to be identical agents, the labor market is composed of many distinct groups that may have different reactions to policy. We compare a bivariate correlated unobserved components model against common filters in the literature to understand the business cycle dynamics of various groups. Our results show that the different filters provide conflicting results for the variability of the series components and the dominant force during recessions. According to our most general model, female unemployment is dominated by the permanent component during recession, while the transitory (or cyclical) component plays a larger role for male unemployment. Therefore, policy enacted to target either structural or cyclical unemployment will have unequal effects in reducing unemployment across genders. These results are robust to different disaggregate specifications and data simulations. Therefore, since males and females have separate and distinctive reactions to macroeconomic shocks, then policy may have unintentionally unequal effects.

C0315: **A time-varying threshold star model of unemployment and the natural rate**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Smooth-transition autoregressive (STAR) models, competitors of Markov-switching models, are limited by an assumed time-invariant threshold level. However, a STAR framework could estimate a time-varying threshold level of unemployment. One can consider this threshold a "tipping level" where the mean and dynamics of the natural rate of unemployment shift. If the threshold level is time-varying, one can add an error-correction term—between the lagged levels of unemployment and the threshold — to the autoregressive terms in the STAR model. Thus, the time-varying latent threshold level serves as both a demarcation between regimes and an error-correction term.

CO116 Room Jessel REGIME CHANGE MODELING I: FINANCE AND FORECAST Chair: Willi Semmler

C0319: **Structural breaks in panel data: Large number of panels and short length time series**
Presenter: **Jan Hanousek**, Charles University, Prague, Czech Republic
Co-authors: Jaromir Antoch, Marie Huskova, Lajos Horvath, Shixuan Wang
The detection of the (structural) break or so called change point problem has drawn increasing attention from both theoretical and applied research over the last decade. A large part of the existing research concentrates on the detection (asymptotic properties) of the change point problem for panels with a large time dimension T. We study a different approach, i.e., we consider the asymptotic properties with respect to N (number of panel members) while keeping T fixed. This case is typically related to large (firm-level) data containing financial information about an immerse number of firms/stocks across a limited number of years/quarters/months. We propose a general approach for testing for the break(s), which also allows their detection. We show the asymptotic behaviour of the test statistics, along with an alternative wild bootstrap procedure that could be used to generate the critical values of the test statistics. The theoretical approach is supplemented by numerous simulations and extended by an empirical illustration. In the practical application we demonstrate the testing procedure in the framework of the four factors CAPM model. We estimate breaks in monthly returns of the US mutual funds during the subprime crises (January 2006 to February 2010).

C0524: **To what extent does sentiment drive real GDP**
Presenter: **Jacek Kotlowski**, Szkola Glowna Handlowa w Warszawie, Poland
Co-authors: Michal Brzoza-Brzezina
Two important observations in the literature are addressed. First, that shocks to expectations about future GDP (e.g. news about technology,

sentiments about future demand) can exert a significant impact on the business cycle. Second, that in structural macroeconomic models it is hard to explain the extent of cyclical co-movement between economies taking into account their trade links only. We check empirically to what extent the observed correlation of business cycles can be explained by spillovers of news/ sentiments from a large (euro area) to a small (Poland) economy. We conduct the analysis using a structural VAR model and identify the shocks by imposing medium run restrictions. Our results show that sentiment demand shocks domestic and foreign explain over 35 percent of the forecast error variance of real GDP in the small economy in a four-year horizon. The other shocks to expectations (news about technology) are significantly less important. Foreign spillovers of expectations (technology and demand) account for almost 20 percent of GDP variance in Poland at all horizons.

C1533: **Arbitrage costs and nonlinear adjustments in individual equity prices: A VSTECM modeling**
Presenter: **Fredj Jawadi**, University of Evry, France

Arbitrage costs cause nonlinear adjustment between equity prices and fundamentals that can be represented by a two-regime ESTECM (Exponential Smooth Transition Error Correction Model). The fundamental values are those estimated previously for 27 of main firms belonging to the French CAC 40 stock price index. These values couple the Dividend Discount Model with the Arbitrage Pricing Theory, the latter determining the long term risk premia included in the discount rate. Accordingly, because fundamental values of equities depend on common factors, the ESTECMs are estimated using a Vector-ESTECM system compounded of the 27 firms. For any firm, it is shown that deviations follow a quasi-random walk in the central regime where prices are near fundamentals (i.e. the mean reversion mechanism is inert when arbitrage costs are greater than expected receipts), while they approach a white noise in the outer regimes (i.e. the mean reversion becomes active when arbitrage cost are lower than expected receipts). As for analyses based on stock price indices, the convergence speed of prices toward fundamentals appeared to depend on the size of the deviation. However, the magnitudes of under- and overvaluation of equity price and the adjustment speed are found to depend strongly both on the date and the firm, which highlight the fact that adjustments of stock price indices hide important disparities between firms.

C1586: **Nonlinear intermediary asset pricing in the oil futures market**
Presenter: **Malte Rieth**, DIW Berlin, Germany
Co-authors: Anton Velinov, Daniel Bierbaumer

The nonlinear trading behavior of financial intermediaries in the oil futures market is studied using structural vector autoregressions with Markov switching in heteroskedasticity. The empirical model is identified through theory and allows for regime-dependent contemporaneous effects and volatility. The results suggest that the demand curve of intermediaries steepens significantly during turbulent times, amplifying the price impact of other traders' demand shocks by two thirds. Moreover, the variance of intermediaries' own demand shocks doubles during these episodes, further raising price volatility. These findings are consistent with the hypothesis that intermediaries' funding constraints are related to volatility and imply nonlinear asset pricing.

CO260 Room Montague FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C0419: **The role of technical indicators in exchange rate forecasting**
Presenter: **Ekaterini Panopoulou**, University of Kent, United Kingdom
Co-authors: Ioannis Souropanis

Forecasting exchange rates is a subject of wide interest to both academics and practitioners. We aim at contributing to this vivid research area by highlighting the role of both technical indicators and macroeconomic predictors in forecasting exchange rates. Employing monthly data ranging from January 1974 to December 2014 for six widely traded currencies, we show that both types of predictors provide valuable information about future currency movements. To efficiently summarise the information content in candidate predictors, we extract the principal components of each group of predictors. Our findings suggest that combining information from both technical indicators and macroeconomic variables significantly improves and stabilises exchange rate forecasts versus using either type of information alone.

C0420: **The normal before the crises: The volatility and skewness crystal ball**
Presenter: **Radu Tunaru**, University of Kent, United Kingdom
Co-authors: Teng Zheng

Merton's jump-diffusion model coupled with Markov Chain Monte Carlo inferential techniques are used to capture the volatility and skewness of the S&P500 index between 1980 and 2015 accounting for parameter estimation risk. We find that the subprime crisis has been preceded a long period when the equity market returns distribution reverted back to a Gaussian distribution indicating a very calm period without extreme events. This event only happened in the U.S. market, where the 2008 crisis originated, and it did not occur in other major economies like Eurozone, the U.K., Germany, Japan or China. Our empirical results are in line with the Minsky theory, and also confirms the theory of endogenous risk. When a calm period is observed, a high level of market sentiment leads to an extra negative impact to the subsequent market returns. Accounting for the interaction between normality and levels of sentiment, this effect may encourage extra risk taking and over expectation of future growth.

C0417: **An examination of herd behaviour in hedge funds**
Presenter: **Nikolaos Voukelatos**, University of Kent, United Kingdom
Co-authors: Ekaterini Panopoulou

The extent to which hedge fund managers tend to herd around a consensus when making investment decisions is examined. To this end, we use the cross-sectional dispersion of hedge fund returns as a measure of aggregate heterogeneity among funds that belong to the same style group. The empirical results suggest that the propensity of fund managers to follow the herd varies across different fund groups, and it is also related to a set of variables at the aggregate market level.

C0480: **Extracting risk neutral densities for weather derivatives pricing using the maximum entropy method**
Presenter: **Antonis Alexandridis**, University of Kent, United Kingdom
Co-authors: Henryk Gzyl, Enrique ter Horst, German Molina

The use of maximum entropy is proposed to extract the risk neutral probabilities directly from the weather market prices. The proposed methodology is computationally fast, model free, non-parametric and can overcome the data sparsity problem that governs the weather market. We infer consistent risk neutral probabilities along with their densities from the market price of temperature options. The risk neutral probabilities inferred from a smaller subset of the data reproduce the other prices and can be used to value accurately all other possible derivatives in the market sharing the same underlying. We examine two sources of the out-of-sample valuation error. First, we use different sets of possible physical state probabilities that correspond to different levels of expertise of the trader. Then, we apply our methodology under three scenarios where the available information in the market is based on historical data, meteorological forecasts or both. Our results indicate that different levels of expertise can affect the accuracy of the valuation. When there is a mix of information, non-coherent sets of prices are observed in the market.

CO126 Room Woburn NETWORK ANALYSIS AND HIGH DIMENSIONAL TIME SERIES MODELS Chair: Christian Brownlees

C0241: **On the architecture of financial networks**
Presenter: **Ruben Hipp**, University of Mannheim, Germany

By using forecast error variance decompositions to identify networks, a popular connectedness measurement has been introduced and a new standard in estimating systemic risk has been successfully established. In contrast, we develop a model to estimate a network within a financial setting without a forecast horizon. By applying a structural vector autoregressive model, it is possible to identify immediate reactions within the system and therefore connections between financial firms. In addition, an application on the U.S. financial market quantifies systemic risk and identifies the most risk receiving and distributing financial firms in a dynamic fashion. Various simulations and applications indicate that a network containing structural VAR allows for a more sophisticated identification of connections and can give insights of how the architecture of financial networks looks like.

C0387: **Community detection in large vector autoregressions**
Presenter: **Gudmundur Gudmundsson**, Universitat Pompeu Fabra, Spain

A family of vector autoregressive models of order 1 (VAR(1)) are introduced where the coefficient matrix is based on an underlying random network. Real-world networks are frequently endowed with a community structure, where the vertices form natural groups within which the frequency of linkages between vertices is higher than without. We therefore focus on the case where the network underlying the model has a community structure and introduce an algorithm to detect the communities consistently. The algorithm is based on spectral clustering and uses the eigenvectors of the coefficient matrix of the VAR(1) for detection. We apply the methodology to study clustering in industrial production and ETF volatility.

C0616: **A multi-factor realized GARCH with an application to the Fama-French model**
Presenter: **Ilya Archakov**, University of Vienna, Austria
Co-authors: Asger Lunde, Peter Hansen

A novel approach is proposed to model and measure systematic risk in equity markets. Asset returns are modeled in a multiple regression framework with GARCH-type dynamics for conditional variances and correlations, which imply temporal variation of the regression coefficients, that are commonly referred to as betas. The model incorporates information from high-frequency based realized measures. These help to identify the latent covariance process and enable the model to promptly adapt to changes. Our framework is consistent with the broad class of linear factor models in the asset pricing literature, and we apply our framework to the famous three-factor Fama-French model in an empirical analysis with more than 800 individual assets. We document an appreciable cross-sectional and temporal variation of the model-implied risk loadings with the especially strong (though short-lived) distortion around the Financial Crisis episode. In addition, we find a significant heterogeneity in a relative explanatory power of the Fama-French factors across the different sectors of economy and detect a fluctuation of the risk premia estimates over time. The empirical evidence emphasizes the importance of taking into account dynamic aspects of the underlying covariance structure in asset pricing models.

C0713: **Mixed-frequency macro-financial spillovers**
Presenter: **Mark Hallam**, City University London, United Kingdom
Co-authors: John Cotter, Kamil Yilmaz

A new methodology is developed to analyse spillovers between the real and financial sides of the economy that employs a mixed-frequency modelling approach. This enables high-frequency financial and low-frequency macroeconomic data series to be employed directly, avoiding the data aggregation and information loss incurred when using common-frequency methods. In a detailed analysis of macro-financial spillovers for the US economy, we find that the additional high-frequency information preserved by our mixed-frequency approach results in estimated spillovers that are typically substantially higher than those from an analogous common-frequency approach and are more consistent with known in-sample events. We also show that financial markets are typically net transmitters of shocks to the real side of the economy, particularly during turbulent market conditions, but that the bond and equity markets act heterogeneously in both transmitting and receiving shocks to the non-financial sector. We observe substantial short and medium-run variation in macro-financial spillovers that is statistically associated with key variables related to financial and macroeconomic fundamentals; the values of the term spread, VIX and unemployment rate in particular appear to be important determinants of macro- financial spillovers.

CO490 Room SH349 MIXTURE MODELS, INDEPENDENT COMPONENTS, AND IDENTIFICATION Chair: Markus Haas

C0671: **Approximating expected shortfall for heavy-tailed distributions**
Presenter: **Jochen Krause**, Valora Management AG, Switzerland
Co-authors: Marc Paolella, Simon Broda

A saddlepoint approximation for evaluating the expected shortfall of financial returns under realistic distributional assumptions is derived. This addresses a need that has arisen after the Basel Committee's proposed move from Value at Risk to expected shortfall as the mandated risk measure in its market risk framework. Unlike earlier results, the approximation does not require the existence of a moment generating function, and is therefore applicable to the heavy-tailed distributions prevalent in finance. A link is established between the proposed approximation and mean-expected shortfall portfolio optimization. Numerical examples include the noncentral t, generalized error, and alpha-stable distributions. A portfolio of DJIA stocks is considered in an empirical application.

C0830: **Financial intermediation and the cross-section of FX returns**
Presenter: **Dennis Umlandt**, Kiel University, Germany
Co-authors: Stefan Reitz

Recent contributions point to the importance of balance sheet variables of financial intermediaries for the cross-section of excess returns in several asset classes. We show that shocks to the equity capital ratio of financial intermediaries explain variations in the returns of carry trades, which are popular investment strategies that borrow in low and invest in high interest rate currencies. To allow for time-varying market prices of risk and betas a regression-based dynamic asset pricing approach is employed. Strong support for time-variation in the market price of risk is found and the equity capital ratio factor performs well in comparison to other popular pricing factors of currency returns.

C0932: **On interdependence and shift contagion between core Euro Area refinancing conditions**
Presenter: **Sebastian Mueller**, Christian-Albrechts-University Kiel, Germany

Changes in linkages between Euro Area (EA 12) country long-term bond yields since the introduction of the Euro in January 2001 are evaluated. These linkages are measured as instantaneous cross market transmissions in a structural vector autoregressive (SVAR) framework. Employing different consistent estimators for scale free linkages we access the qualitative and quantitative robustness of statistical identification in small samples based on unconditional as well as on conditional patterns of heteroscedasticity. We document interdependence, shifts and break-ups in structural relationships of core EA country government bond yields. Moreover, the relative importance of country specific news for the refinancing

conditions of others are revealed. We test for changes in pricing rules in the Euro crisis and find the ECB announcement to buy and mutualize government debt was only partially effective in reestablishing pre crisis interdependencies.

C1204: **Robust and flexible mixture models for the identification of structural shocks of financial time series**
Presenter: **Markus Haas**, University of Kiel, Germany
Co-authors: Sebastian Mueller
Dynamic mixture or regime-switching models have long been popular in empirical finance due to their good fit and economic interpretability of the extracted regimes. More recently, these models have been employed to identify the contemporaneous structural effects in vector autoregressive models. Gaussian mixtures are typically used in this context, with all structural shock processes changing regimes simultaneously. However, the assumption of (simultaneous) Gaussian regimes is often inappropriate for the structural shocks driving multivariate financial time series, and this may seriously distort the identification and economic interpretation of these shocks. Thus, currently existing approaches to identification via regime-switching effects are extended in order to accommodate thick-tailed innovations and independently switching components. The usefulness of the models is illustrated by applying them to a set of relevant problems in financial economics, such as transmission of shocks in the euro area and price formation in foreign exchange markets.

CG072 Room Gordon CONTRIBUTIONS IN STOCHASTIC VOLATILITY Chair: Peter Exterkate

C1517: **Semi-parametric Bayesian forecasting with an application to stochastic volatility**
Presenter: **Martina Danielova Zaharieva**, University of Muenster, Germany
Co-authors: Fabian Goessling
A new and highly flexible Bayesian sampling algorithm is proposed for non-linear state space models under non-parametric 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 constraining the models by transformations or the need for conjugate distributions. We use the Chinese restaurant representation of the Dirichlet process mixture, which allows for a parsimonious and generally applicable sampling algorithm. Thus, our estimation algorithm combines a pseudo marginal Metropolis Hastings scheme with a marginalized hierarchical semi-parametric model. We test our approach for several nested model specifications using simulated data and provide density forecasts. Furthermore, we carry out a real data example.

C1458: **A macro-finance term structure model with volatility-induced stationarity**
Presenter: **Anne Hansen**, University of Copenhagen and Danmarks Nationalbank, Denmark
Term structure models with a particular form of stochastic volatility in which mean-reversion is induced through level-dependent conditional volatility are studied. In a simulation study, it is shown that the celebrated Gaussian affine term structure model (GATSM) produces biased term premia estimates if the data is generated by processes that exhibit VIS. The relevance of this finding is established in context of US Treasury yields by showing that a macro-finance term structure model that accounts for VIS outperforms the GATSM in terms of (i) matching state vector dynamics both in- and out-of-sample; (ii) obtaining realistic term premia estimates; and (iii) match survey forecasts of future short rates and inflation rates. Moreover, the term structure model that accounts for VIS have economically sensible implications, whereas the GATSM encompasses several puzzles. These findings guide the term structure specification for monetary policy makers and debt managers for whom term premia estimates are of particular interest.

C0236: **Volume, durations and jumps in SV models for the evolution of intraday financial volatility**
Presenter: **Antonio Santos**, University of Coimbra, Portugal
The use of stochastic volatility (SV) models in the characterisation of the intraday volatility evolution is analyzed. The novelty is the introduction of a set of information elements that can be retrieved from intraday data, like volume of transactions and durations. This kind of information allows also the consideration of jumps in returns and respective volatility, which is done through a subordinated duration stochastic process. A bivariate state-space model relates returns and durations with a latent state. The probabilities of the jumps are modelled through a process using a logit kind of relationship, which allows time-varying probabilities for the jumps. Volume-domain returns are used, which when mapped to the time-domain, represent unequal time-spaced observations, and the duration between the successive volume-domain returns can be calculated. This kind of variable may give information on the evolution of the volatility defined through a state-space model, and can be related to the probability of a jump in returns or volatility. Our approach is tested using 2 years of intraday data associated with 10 stocks traded in US markets.

C0562: **Bayesian state-space model with time varying parameters and stochastic volatility in identification of financial shocks**
Presenter: **Seyma Vahap**, University of Strathclyde, United Kingdom
The aim is to examine the nature and evolving features of exogenous shocks to credit spreads as a source of business cycle fluctuations by allowing for a dynamic feedback between macroeconomics and finance within a Bayesian time varying parameter structural vector autoregressive (SVAR) model with stochastic volatility. We identify an unanticipated increase in credit spread that can be considered exogenous with respect to both macroeconomic and financial variables using time and sign restrictions. We find that these shocks have contributed to significantly substantial contraction is economic activity. However, the response of short rates remains modest. The results also show that innovations to credit spreads cause ambiguity in the response of inflation and stock returns.

Saturday 16.12.2017 14:35 - 16:15 Parallel Session E – CFE-CMStatistics

EI736 Room CLO B01 STATISTICAL INFERENCE AND MACHINE LEARNING Chair: Jelena Bradic

E0466: **Learning structured densities without parametric assumptions**
Presenter: **Mladen Kolar**, University of Chicago, United States
Learning the structure of a probabilistic graphical models is a well studied problem in the machine learning community due to its importance in many applications. Current approaches are mainly focused on learning the structure under restrictive parametric assumptions,which limits the applicability of these methods. We study the problem of estimating the structure of a probabilistic graphical model without assuming a particular parametric model. We consider probabilities that are members of an infinite dimensional exponential family. One difficulty in learning nonparametric densities is evaluation of the normalizing constant. In order to avoid this issue, our procedure minimizes the penalized score matching objective. We show how to efficiently minimize the proposed objective using existing group lasso solvers. Furthermore, we prove that our procedure recovers the graph structure with high-probability under mild conditions. Simulation studies illustrate ability of our procedure to recover the true graph structure without the knowledge of the data generating process.

E0468: **Efficient policy learning**
Presenter: **Stefan Wager**, Stanford University, United States
Co-authors: Susan Athey
There has been considerable interest across several fields in methods that reduce the problem of learning good treatment assignment policies to the problem of accurate policy evaluation. Given a class of candidate policies, these methods first effectively evaluate each policy individually, and then learn a policy by optimizing the estimated value function; such approaches are guaranteed to be risk-consistent whenever the policy value estimates are uniformly consistent. However, despite the wealth of proposed methods, the literature remains largely silent on questions of statistical efficiency: there are only limited results characterizing which policy evaluation strategies lead to better learned policies than others, or what the optimal policy evaluation strategies are. We build on classical results in semiparametric efficiency theory to develop quasi-optimal methods for policy learning; in particular, we propose a class of policy value estimators that, when optimized, yield regret bounds for the learned policy that scale with the semiparametric efficient variance for policy evaluation. On a practical level, our result suggests new methods for policy learning motivated by semiparametric efficiency theory.

EO019 Room MAL B18 GRAPHICAL MARKOV MODELS II Chair: Nanny Wermuth

E1243: **Identical maximum likelihood estimates for Gaussian and Ising models defined by a chordless cycle**
Presenter: **Giovanni Marchetti**, University of Florence, Italy
Undirected graphical models defined by a chordless cycle require in general an iterative fitting procedures to get maximum likelihood estimates. For Gaussian models, the canonical parameters are the concentrations, that is the off-diagonal element in the inverse covariance matrix, while for Ising models, are the log-linear, two-factor interactions. However, we show conditions under which, if the canonical parameters are transformed to partial correlations, the two different likelihood functions, one for the continuous and the other for the binary variables, give the same maximum likelihood estimates provided the relevant starting correlation matrices coincide and have a closed form.

E0635: **Probability based independence sampler for Bayesian quantitative learning in graphical log-linear marginal models**
Presenter: **Claudia Tarantola**, University of Pavia, Italy
Co-authors: Monia Lupparelli, Ioannis Ntzoufras
A fully automatic and efficient MCMC strategy is presented for quantitative learning for graphical log-linear marginal models. While the prior is expressed in terms of the marginal log-linear parameters, we build an MCMC algorithm which employs a proposal on the probability parameter space. The corresponding proposal on the marginal log-linear interactions is obtained via parameter transformations. By this strategy, we achieve to move within the desired target space. At each step we directly work with well-defined probability distributions. Moreover, we can exploit a conditional conjugate setup to build an efficient proposal on probability parameters. The proposed methodology is illustrated using a popular four-way dataset.

E1281: **Permutation-based causal inference algorithms with interventions**
Presenter: **Caroline Uhler**, Massachusetts Institute of Technology, United States
Co-authors: Yuhao Wang, Liam Solus, Karren Yang
A recent break-through in genomics makes it possible to perform perturbation experiments at a very large scale. In order to learn gene regulatory networks from the resulting data, efficient and reliable causal inference algorithms are needed that can make use of both, observational and interventional data. We will present the first provably consistent such algorithm. It is a hybrid approach that uses conditional independence relations in a score-based method. Hence, this algorithm is non-parametric, which makes it useful for analyzing inherently non-Gaussian gene expression data. We will end by analyzing its performance on simulated data, protein signaling data, and single-cell gene expression data.

E0894: **Maximum likelihood estimation of the latent class model through model boundary decomposition**
Presenter: **Piotr Zwiernik**, Universitat Pompeu Fabra, Spain
The Expectation-Maximization (EM) algorithm is routinely used for the maximum likelihood estimation in the latent class analysis. However, the EM algorithm comes with no guarantees of reaching the global optimum. We study the geometry of the latent class model in order to understand the behavior of the maximum likelihood estimator. In particular, we characterize the boundary stratification of the binary latent class model with a binary hidden variable. For small models, such as for three binary observed variables, we show that this stratification allows exact computation of the maximum likelihood estimator. In this case we use simulations to study the performance of the EM algorithm.

EO200 Room MAL B20 EXTREME VALUE STATISTICS Chair: John Einmahl

E0281: **Tail dimension reduction for extreme quantile estimation**
Presenter: **Laurent Gardes**, University of Strasbourg, France
In a regression context where a real response variable Y is recorded with a p-dimensional covariate X, two situations can occur simultaneously: (a) we are interested in the tail of the conditional distribution and not on the central part of the distribution and (b) the number p of regressors is large. Up to our knowledge, these two situations have only been considered separately in the literature. The aim is to propose a new dimension reduction approach adapted to the tail of the distribution in order to propose an efficient conditional extreme quantile estimator when the dimension p is large. The results are illustrated on simulated data and on a real dataset.

E0499: **Risk measure estimation for $\beta-$mixing time series and applications**
Presenter: **Armelle Guillou**, Strasbourg, France
Co-authors: Valerie Chavez-Demoulin
The application of extreme-value theory in the context of stationary $\beta-$mixing sequences that belong to the Frechet domain of attraction is discussed. In particular, we propose a methodology to construct bias-corrected tail estimators. Our approach is based on the combination of two estimators of the extreme-value index to cancel the bias. Then the resulting estimator is used to estimate an extreme quantile. In a simulation study, we outline the performance of our proposals that we compare to alternative estimators recently introduced in the literature. Also, we compute the asymptotic variance in specific examples when possible. Our methodology is applied to two datasets on finance and environment.

E0577: **Extreme value estimation for censored regularly varying tails**
Presenter: **Jan Beirlant**, KULeuven, Belgium
Motivated by applications in long-tail insurance products, we consider tail estimation for right censored data from regularly varying tail models. The bias of the available estimators of the extreme value index can be substantial and depends strongly on the amount of censoring. We review the existing literature, propose a new bias reduced estimator, and show how shrinkage estimation can help to keep the MSE under control. Basic asymptotics results are provided, and a bootstrap algorithm is proposed to construct confidence intervals. We compare these new proposals with the existing estimators through simulation. We also consider the corresponding bivariate problem. Throughout we consider a motor third party liability case from a European country.

E0942: **Tail behaviour of a multivariate quantile based on optimal transport**
Presenter: **Cees de Valk**, Universita catholique de Louvain, Belgium
Co-authors: Johan Segers
Recently, an attractive multivariate quantile based on Monge-Kantorovitch optimal transport was proposed as a refinement of half-space depth: a map transforming a spherical reference distribution function to the distribution function of interest is sought which minimises the mean square of the size of the displacement. We explore the use of this idea for estimation of multivariate tail quantiles. Beginning with the assumption of multivariate regular variation, we consider limit relations satisfied by the optimal map and the associated quantile. An example of an application is discussed.

EO230 Room MAL B30 FUNCTIONAL DATA ANALYSIS Chair: Alicia Nieto-Reyes

E0702: **Functional logistic regression: An RKHS approach**
Presenter: **Jose Berrendero**, Universidad Autonoma de Madrid, Spain
Co-authors: Beatriz Bueno-Larraz, Antonio Cuevas
A functional logistic regression model is proposed to explore the relationship between a dichotomous response variable and a functional predictor. The proposal is based on ideas borrowed from the theory of reproducing kernel Hilbert spaces (RKHS). Similarly to the finite-dimensional case, our model holds when the conditional distributions of the predictor given the two possible values of the response are Gaussian with the same covariance structure. Moreover, some particular choices of the slope function lead to the point-impact model. We also give conditions (which include Brownian-like predictors) under which the maximum likelihood estimator of the slope function does not exist with probability one and address some possible solutions.

E0759: **Robust estimators under a functional partial linear model**
Presenter: **Graciela Boente**, Universidad de Buenos Aires, Argentina
Co-authors: Matias Salibian-Barrera, Pablo Vena
Functional data analysis provides modern analytical tools for data that are recoded as images or as a continuous phenomenon over a period of time. Because of the intrinsic nature of these data, they can be viewed as realizations of random functions often assumed to be in a Hilbert space such as L2(I), with I a real interval or a finite dimensional Euclidean set. Partial linear modelling ideas have recently been adapted to situations in which functional data are observed. More precisely, two generalizations have been considered to deal with the problem of predicting a real-valued response variable using explanatory variables that include a functional element, usually a random function, and a random variable. The semi-functional partial linear regression model allows the functional explanatory variables to act in a free nonparametric manner, while the scalar covariate corresponds to the linear component. On the other hand, the so-called functional partial linear model assumes that the scalar response is explained by a linear operator of a random function and a nonparametric function of a real–valued random variable. We will briefly discuss some approaches leading to obtain robust estimators functional partial linear model.

E0906: **Prediction bands for functional data based on depth measures**
Presenter: **Antonio Elias**, Universidad Carlos III de Madrid, Spain
A new methodology is proposed for predicting a partially observed curve from a functional data sample. The novelty of our approach relies on the selection of sample curves which form tight bands that preserve the shape of the curve to predict, making this a deep datum. The involved subsampling problem is dealt by two algorithms specially designed to be used in conjunction with two different ways for computing central regions for visualizing functional data. From this merge we obtain prediction bands for the unobserved part of the curve in question. We test our algorithms by forecasting the Spanish electricity demand and imputing missing daily temperatures. The results are consistent with our simulation that show that we are able to predict at the far horizon.

E1067: **Joint diagonalisation of scatter operators: Functional fourth order blind identification**
Presenter: **Germain Van Bever**, Universite libre de Bruxelles, Belgium
Co-authors: Hannu Oja, Frank Critchley, Radka Sabolova, Bing Li
With the increase in measurement precision, functional data is becoming common practice. Relatively few techniques for analysing such data have been developed, however, and a first step often consists in reducing the dimension via Functional PCA, which amounts to diagonalising the covariance operator. Joint diagonalisation of a pair of scatter functionals has proved useful in many different setups, such as Independent Component Analysis (ICA), Invariant Coordinate Selection (ICS), etc. The main aim consists in extending the Fourth Order Blind Identification procedure to the case of data on a separable Hilbert space (with classical FDA setting being the go-to example). In the finite-dimensional setup, this procedure provides a matrix G such that GX has independent components, if one assumes that the random vector X satisfies $X = PZ$, where Z has independent marginals and P is an invertible mixing matrix. When dealing with distributions on Hilbert spaces, two major problems arise: (i) the notion of marginals is not naturally defined and (ii) the covariance operator is, in general, non invertible. These limitations are tackled by reformulating the problem in a coordinate-free manner and by imposing natural restrictions on the mixing model. The proposed procedure is shown to be Fisher consistent and affine invariant. A sample estimator is provided and its convergence rates are derived. The procedure is amply illustrated on simulated and real datasets.

EO510 Room MAL B33 MODELING DEPENDENCE FOR FUNCTIONAL DATA Chair: Matthew Reimherr

E1536: **Functional GARCH models**
Presenter: **Clement Cerovecki**, Univ libre de Bruxelles, Belgium
Co-authors: Siegfried Hormann, Christian Francq, Jean-Michel Zakoian
Increasing availability of high frequency data has triggered many new research areas in statistics. Functional data analysis (FDA) is one of these disciplines. In FDA densely observed data are transformed into curves and then each (random) curve is considered as one data object. A natural, but still relatively unexplored context for FDA methods is related to financial data, where high-frequency trading nowadays takes a significant proportion of trading volumes. Recently, articles on functional versions of the famous ARCH and GARCH models have been brought fourth. Due to their technical complexity, existing estimators of the underlying functional parameters are moment based, an approach which is known to be relatively inefficient in this context. We promote quasi likelihood approaches. We focus on a finite dimensional and hence feasible framework which allows a realistic practical implementation. Moreover, we can go beyond consistency results and are able to obtain asymptotic normality of the estimators. We support the superiority of our approach by simulations and illustrate its use by forecasting realized volatility of the S&P100 market index.

E1500: **A moment-based notion of time dependence for functional time series**
Presenter: **Nazarii Salish**, University Carlos III de Madrid, Spain
The focus is on the fundamental topic of time dependence for time series when data points are given as functions. We construct a notion of time dependence through the scores of the principal components, which allows us to adapt various scalar time series techniques to the functional data context. In particular, we define dependence based on the autocovariances and cumulants of the scores, covering short and long memory scenarios. We show that this notion naturally applies to linear processes. To justify this moment based approach we investigate the asymptotic properties of the estimator of the functional principal components and show its consistency under short and long memory. Finally, applicability of our notion is illustrated through several statistical problems in functional time series: estimation of the functional autoregressive model, estimation of the long-run covariance function and testing for short memory against the long memory alternative.

E1506: **Testing separability of functional time series**
Presenter: **Panayiotis Constantinou**, The Pennsylvania State University, United States
Co-authors: Piotr Kokoszka, Matthew Reimherr
A significance test is derived and studied for determining if a panel of functional time series is separable. In this context, separability means that the covariance structure factors into the product of two functions, one depending only on time and the other depending only on the coordinates of the panel. Separability is a property which can dramatically improve computational efficiency by substantially reducing model complexity. It is especially useful for functional data as it implies that the functional principal components are the same for each member of the panel. However such an assumption must be verified before proceeding with further inference. Our approach is based on functional norm differences and provides a test with well controlled size and high power. In addition to an asymptotic justification, our methodology is validated by a simulation study. It is applied to functional panels of particulate pollution and stock market data.

E1588: **Generalized functional linear models with points of impact**
Presenter: **Dominik Poss**, University of Bonn, Germany
Co-authors: Dominik Liebl
A generalized functional linear regression model with points of impact is assumed. In the classical generalized functional linear regression model, scalar responses $Y_1,\dots,Y_n$ are connected to the inner product of functional predictors $X_1,\dots,X_n$ and an unknown coefficient function via a smooth link function. Additionally, an unknown number of unknown specific locations ("points of impact") at which the functional predictor will have a further effect on the response are allowed. The focus is on the estimation of these points of impact. Some theoretical results are given and the estimation procedure is illustrated in the case of a (functional) logistic regression framework with points of impact where the depend variable Y_i is binary.

EO284 Room MAL B34 STATISTICS FOR HILBERT SPACES Chair: Gil Gonzalez-Rodriguez

E0558: **Robust estimation of the total electricity load curve by sampling in a finite population**
Presenter: **Camelia Goga**, Universite de Bourgogne, France
Many studies carried out in the French electricity company (EDF) are based on the analysis of total electricity consumption curves measured at a small time scale (often half-hourly) for one or several groups of clients sharing some common characteristics (e.g. customers of the same electricity provider, having a particular electric equipment or living in a given geographic area). The aim of these studies can be, for example, to assist the power grid manager in maintaining the balance between electricity consumption and production at every instant on the power grid. The total consumption curves can also be used to help the Sales Division to quantify the impact of a specific electric use or equipment on the electricity consumption, to build new innovative pricing strategies or to create new services based on customers consumption analysis.

E1071: **Depth-based methods for sparse and complex functional data**
Presenter: **Sara Lopez Pintado**, Columbia University, United States
Data depth is a well-known and useful notion in functional data analysis. It provides a center-outward ranking for a sample of curves. This ordering allows the definition of descriptive statistics such as medians, trimmed means and central regions for functional data. Moreover, data depth is often used as a building block for developing outlier-detection techniques and for robustifying standard statistical methods. We consider complex functional data such as images and we introduce depth-based location and dispersion measures. Permutation test for comparing location and dispersion of two groups of images are proposed and calibrated. In addition, techniques for detecting image outliers are introduced. The performances of these methods are illustrated in simulated and real data sets. We have also extended the notion of depth to sparse functional data where the functions are observed in subject dependent and/or sparse grids. In this case the functional data is an estimate of the underlying true curves and there is uncertainty in its estimation. We propose a notion of depth that takes into account this uncertainty.

E1294: **Application of functional correlation in biology and econometrics**
Presenter: **Christian Ritz**, University of Copenhagen, Denmark
Co-authors: Anne van Delft
In much applied, exploratory research correlation coefficients are used as a means for gauging the strength of relationships between different outcomes or responses measured or recorded on the same entities, subject, or units, repeatedly over time. These coefficients may be indicative of presence or absence of an association or linkage between underlying mechanisms or processes, e.g., test of independence. Both within biology and econometrics data are typically unbalanced, either rich or sparse, and it remains challenging to provide describe such data in terms of standard statistical models. For instance, standard Pearson or Spearman correlation coefficients may be calculated between pairs of responses for each time point or for all time points pooled. In both cases dependencies over time are not taken into account. However, these data fit perfectly within the

framework of functional data and functional data methods seem well suited for analyzing such data. Recently, a functional correlation coefficient was proposed using singular value decompositions. The aim of the present study is to further demonstrate the usefulness of the functional correlation coefficient. To our knowledge the concept of a functional correlation coefficient has not so far been utilized in biology and econometrics in any noticeable extent.

E1134: **Semi-supervised classification for functional data and its applications**
Presenter: **Yoshikazu Terada**, Osaka University; RIKEN, Japan
In various fields, data recorded continuously during a time interval and curve data such as spectral data become common. These kinds of data can be interpreted as "functional data." We have studied binary semi-supervised classification problem for functional data. For example, in the sports medicine field, it is important to identify players who are at-risk for career-threatening injuries based on the various functional data reflecting individual motor dynamics. For this problem, the usual supervised classification methods are not appropriate since it is not necessary that all the at-risk players will have serious injury during the experimental period. In this situation, we consider binary classification problem from only positive and unlabeled functional data. We propose a simple classification algorithm for this problem. In addition, we prove that, under mild regularity conditions similar to those in a supervised context, the proposed algorithm can achieve perfect asymptotic classification in the context of PU classification. In fact, we show that the proposed algorithm works well not only in numerical experiments but also for real data examples.

EO142 Room MAL B35 STATISTICAL ADVANCES IN NEUROIMAGING Chair: Timothy Johnson

E0322: **A time-varying AR coefficient model of functional near-infrared spectroscopy data**
Presenter: **Timothy Johnson**, University of Michigan, United States
Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low cost, portable, and non-invasive method to monitor brain activity. Similar to fMRI, it measures changes in the level of blood oxygen in the brain. Its time resolution is much finer than fMRI, however its spatial resolution is much courser—similar to EEG or MEG. fNIRS is finding widespread use on young children whom cannot remain still in the MRI magnet and it can be used in situations where fMRI is contraindicated—such as with patients whom have cochlear implants. We propose a fully Bayesian time-varying autoregressive model to analyze fNIRS data. The hemodynamic response function is modeled with the canonical HRF and the low frequency drift with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive process vary with time. Via a simulation studies, we show that this model naturally handles motion artifacts and gives good statistical properties. The model is then apply to a fNIRS study.

E0451: **Semiparametric estimation under shape invariance for fMRI data**
Presenter: **Nicole Lazar**, University of Georgia, United States
Co-authors: Cheolwoo Park, Christopher Helms
The aim is to introduce a semiparametric functional data analysis approach under shape invariance for group comparisons in functional magnetic resonance imaging (fMRI) data. The components of this analysis suite include: function estimation using local polynomial regression; a shape invariant model for the relevant function estimates; evolutionary algorithms. The approach will be demonstrated on a study of practice effects.

E0637: **Principal directions of mediation**
Presenter: **Martin Lindquist**, Johns Hopkins University, United States
Mediation analysis is an important tool in the behavioral sciences for investigating the role of intermediate variables that lie in the path between a treatment and an outcome variable. The influence of the intermediate variable on the outcome is often explored using a linear structural equation model (LSEM), with model coefficients interpreted as possible effects. While there has been significant research on the topic, little work has been done when the intermediate variable (mediator) is a high-dimensional vector. We introduce a novel method for identifying potential mediators in this setting called the principal directions of mediation (PDMs). PDMs linearly combine potential mediators into a smaller number of orthogonal components, with components ranked based on the proportion of the mediating effect each accounts for. We demonstrate the method using a functional magnetic resonance imaging (fMRI) study of thermal pain where we are interested in determining which brain locations mediate the relationship between the application of a thermal stimulus and self-reported pain.

E1202: **Calculating a generated effect modifier (GEM) for treatment selection based on imaging data**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Hyung Park, Eva Petkova, Thaddeus Tarpey
A major goal in precision medicine is to make optimal patient-specific treatment decisions using data observed at baseline. For the treatment of neuropsychiatric disorders, available data may include clinical variables and measures of bevavioral/cognitive performance, as well as complex imaging data. We will present methods for (1) determining low-dimensional projections of all these data that are useful for describing differential treatment response; and (2) for estimating nonparametrically defined link functions based on these projections for each potential treatment. The resulting model can potentially provide powerful tools for precision medicine.

EO067 Room Bloomsbury RECENT ADVANCES IN STATISTICAL GENETICS Chair: Florian Frommlet

E1142: **Mixed model approach for QTL mapping in inbred crosses**
Presenter: **Malgorzata Bogdan**, University of Wroclaw, Poland
Co-authors: Piotr Szulc, David Siegmund, Rebecca Doerge
The problem of mapping of quantitative trait loci (QTL) in experimental crosses in the presence of polygenic effects will be discussed. Specifically, we will explain the phenomenon of so called ghost QTL and how it is related to the problem of hot-spots in expression QTL studies. We will also present a novel methodology based on mixed models which eliminates ghost QTL and allows for efficient identification of true QTL in the presence of polygenic effects

E1284: **Independence hypothesis weighting in biostatistical practice**
Presenter: **Dominic Edelmann**, DKFZ Heidelberg, Germany
Co-authors: Axel Benner
Testing in molecular data often involves a very a large number of hypotheses. In these scenarios, controlling the false discovery rate via Benjamini-Hochberg adjustment of the p-value can lead to very few rejections. To increase the number of discoveries, practitioners commonly apply an unsupervised filtering procedure concentrating the analysis on the most promising hypotheses. However, this filtering procedure is often used quite arbitrarily. Recently, a data-driven technique has been developed, which automatically selects weights for multiple hypothesis testing using side information. Optimizing a target criterion based on the number of rejections, these techniques can dramatically increase the number of discoveries. Beside its complexity compared to unsupervised filtering, one of the main reasons for the disregard of these approaches is that many researchers are not aware that they have important side information available. We present the Independent Hypothesis Weighting (IHW) method and other recently developed data-driven techniques. Moreover, we demonstrate that useful side information is nearly always available for molecular data. Potential

covariates include scale and location parameters, gene annotation and text-mining data. We show that this information considerably increases the number of discoveries in large-scale genetic applications.

E1173: **Statistical tests for genomic time series data**
Presenter: **Andreas Futschik**, JKU Linz, Austria
Co-authors: Sonja Zehetmayer

Genomic time series data arise for instance in experimental evolution, where several populations of fast reproducing organisms, such as yeast or Drosophila, are kept under different environmental conditions. The goal is to find genomic signatures of adaptive selection. Recent experiments revealed however, that different experimental populations do not necessarily show the same genomic signatures of adaptation. This may be for instance due to the random loss of beneficial variants due to genetic drift, due to initial linkage with deleterious alleles, or to different possibilities for adaptation. We investigate different possibilities for identifying selection under such circumstances, and propose new more powerful tests of a global null hypothesis.

E1795: **Outlier detection with mixtures of Gaussian and heavy-tailed distributions**
Presenter: **Alexandra Posekany**, Danube University Krems, Austria

Linear models generally assume normality, a prerequisite often disregarded by data in fields like biology or economics. Our primary aim is to robustify Bayesian inference with mixture models which simultaneously allows for density estimation and outlier detection. To this end, we suggest mixing Students t distributed components in addition to Gaussian ones for identifying the over-dispersed part of data a part of which is extremely noisy, while the rest is normally distributed. To this effect, we employ microarray data as a case study for this behaviour, as they are well-known for their complicated, over-dispersed noise behaviour. Our secondary goal is to present a methodology, which helps not only to identify noisy genes but also to recognise whether single arrays are responsible for this behaviour.

EO368 Room Chancellor's Hall NON-STATIONARITY AND HIGH-DIMENSIONALITY IN TIME SERIES ANALYSIS Chair: Piotr Fryzlewicz

E0191: **Detecting granular time series in large panels**
Presenter: **Christian Brownlees**, UPF, Spain
Co-authors: Geert Mesters

Large economic and financial panels often contain time series that influence the entire cross-section. We name such series granular. We introduce a panel data model that allows us to formalize the notion of granular time series. We then propose a methodology, which is inspired by the network literature in statistics and econometrics, to detect the set of granulars when such set is unknown. The influence of the ith series in the panel is measured by the norm of the ith column of the inverse covariance matrix. We show that a detection procedure based on the column norms allows us to consistently select granular series when the cross-section and time series dimensions are sufficiently large. Moreover, the asymptotic distribution of the column norms is derived in order to construct confidence intervals and carry out hypothesis testing. Importantly, we show that the methodology allows to consistently detect granulars also when the series in the panel are influenced by common factors. A simulation study shows that the proposed procedures perform satisfactorily in finite samples. We illustrate the methodology with applications in macroeconomics and finance.

E0700: **Integrating regularized covariance matrix estimators**
Presenter: **Phoenix Feng**, London School of Economics, United Kingdom
Co-authors: Clifford Lam

When the dimension p is close to or even larger than the sample size n, regularizing the sample covariance matrix is the key to obtaining a satisfactory covariance matrix estimator. One branch of regularization assumes specific structures for the true covariance matrix Σ_0. Another one regularizes on the eigenvalues directly without structure assumptions. The former makes sense when one is confident in a specific structure, while the latter is sound when no specific structures are known. Under more a practical scenario where one is not certain of which regularization method to use, we introduce an integration covariance matrix estimator which is a linear combination of a rotation-equivariant estimator and a regularized covariance matrix estimator assuming a specific structure for Σ_0. We estimate the weights in the linear combination and show that they asymptotically go to the true underlying weights. To generalize, we can put two regularized estimators into the linear combination, each assumes a specific structure for Σ_0. We demonstrate the superior performance of our estimator when compared to other state-of-the-art estimators through extensive simulation studies and real data analyses.

E1019: **Network time series**
Presenter: **Kathryn Leeming**, University of Bristol, United Kingdom
Co-authors: Marina Knight, Guy Nason, Matthew Nunes

A network time series describes observations collected at nodes on a network over time. This network may be known, or require construction from the time series or other information, such as location. The NARIMA (Network ARIMA) framework will be introduced which, although similar to VARMA modelling, allows for changes in the network structure over time. This framework allows modelling and forecasting of large data-sets using few model parameters. As in univariate time series modelling, it is important to take account of any changes in the trend of the network time series. Methods for removing trend will be discussed and demonstrated using an epidemiological example.

E0898: **Positive-definite wavelet estimation of time-varying spectral density matrices**
Presenter: **Rainer von Sachs**, Universita catholique de Louvain, Belgium
Co-authors: Joris Chau

In nonparametric estimation of the spectral density matrix of a multivariate time series it is important to preserve positive-definiteness of the estimator. To this purpose, in previous work the authors have considered multivariate spectral estimation on the Riemannian manifold of Hermitian and positive-definite matrices based on a geometric wavelet approach. Nonlinear wavelet curve denoising on the Riemannian manifold allows one to capture local smoothness behaviour of the spectral matrix across frequency, but also varying degrees of smoothness across components of the spectral matrix. We discuss extensions of this approach to a) non-stationary, i.e. time-varying analyses (the underlying spectral density is allowed to change over time) and b) to situations of replicated multivariate time series as treated in the functional mixed-effects model approach of the authors.

EO382 Room Court RECENT ADVANCES ON HIGH-DIMENSIONAL STATISTICS Chair: Ines Wilms

E0529: **Penalized estimation of sparse high-dimensional single-equation error correction models**
Presenter: **Etienne Wijler**, Maastricht University, Netherlands
Co-authors: Stephan Smeekes
The estimation of high-dimensional single-equation error correction models is considered. We adopt an l1-penalty to obtain sparse parameter estimates and we consider encouraging additional sparsity through the use of a group penalty on the error correction term. Oracle properties are derived under relatively mild assumptions on the error term while allowing the number of parameters to diverge along with the time dimension. Under the strict assumption of Weak Exogeneity (WE) the estimator can be interpreted as an alternative to a well-known test of cointegration. However, without the assumption of WE, our estimation procedure consistently estimates a set of pseudo-parameters for which we obtain analytical expressions. The predictive capability of our method is demonstrated through the use of simulations as well as an empirical application.

E0729: **Volatility spillovers and heavy tails: A large t-vector autoregressive approach**
Presenter: **Luca Barbaglia**, KU Leuven, Belgium
Co-authors: Christophe Croux, Ines Wilms
Volatility is a key measure of risk in financial analysis. The high volatility of one financial asset today could affect the volatility of another asset tomorrow, with important implications for portfolio management. These lagged effects among volatilities - which we call volatility spillovers - are studied using the Vector AutoRegressive (VAR) model. We account for the possible fat-tailed distribution of the VAR model errors using a VAR model with errors following a multivariate Student t distribution with unknown degrees of freedom. Moreover, we study volatility spillovers among a large number of assets. To this end, we use penalized estimation of the VAR model with t-distributed errors. We study volatility spillovers among a large number of energy, biofuel and agricultural commodities. Using network analysis, we reveal bidirectional volatility spillovers between energy and biofuel, and between energy and agriculture commodities.

E0596: **Assessing selection stability in regularised regression models**
Presenter: **Garth Tarr**, University of Sydney, Australia
The mplot R package provides an implementation of model stability and variable inclusion plots for researchers to use to better inform the variable selection process. The initial focus was on exhaustive searches through the model space, however, this quickly becomes infeasible for high dimensional models. An alternative approach for high dimensional models is to combine bootstrap model selection with regularisation procedures. There exists a number of fast and efficient method regularisation methods for variable selection in high dimensional regression settings. We have implemented variable inclusion plots and model stability plots using the glmnet package. We demonstrate the utility of the mplot package in identifying stable regularised model selection choices with respect to two main sources of uncertainty. Firstly, by resampling the data we are able to determine how often various models are chosen when the data changes. Secondly, we are able to evaluate how often competing models are chosen across a range of values for the tuning parameter. Exploring these two sources of uncertainty in model selection generates a large amount of raw data that needs to be processed. The mplot package provides a variety of methods to visualise this raw data to help inform a researcher's model selection choice.

E1133: **Estimation of sparse vector autoregressive moving averages**
Presenter: **Ines Wilms**, KU Leuven; Cornell University, Belgium
Co-authors: Jacob Bien, David Matteson, Sumanta Basu
The Vector AutoRegressive Moving Average (VARMA) model is a fundamental tool for modeling multivariate time series. Recently, a growing interest has arisen in high-dimensional models, where the number of marginal time series is increasingly large. However, as the number of time series increases, the VARMA model becomes heavily overparameterized. For such high-dimensional VARMA models, estimation is generally intractable. In this setting, the high-dimensional Vector AutoRegression (VAR) model has been favored, in both theory and practice. We propose adapting modern regularization methods to estimate high-dimensional VARMA models. Our estimation method is sparse, meaning many model parameters are estimated as exactly zero. The proposed framework has good estimation and forecast accuracy under numerous simulation settings. We illustrate the forecast performance of the sparse VARMA models for several application domains, including macro-economic forecasting, demand forecasting and volatility forecasting. The sparse VARMA estimator gives parsimonious forecast models that lead to important gains in relative forecast accuracy.

EO198 Room G11 NONPARAMETRIC MODELLING OF DEPENDENT DATA Chair: Tatyana Krivobokova

E1844: **Testing for trends in high-dimensional time series**
Presenter: **Wei Biao Wu**, University of Chicago, United States
Statistical inference for trends of high-dimensional time series is considered. Based on a modified L2-distance between parametric and nonparametric trend estimators, we propose a de-diagonalized quadratic form test statistic for testing patterns on trends, such as linear, quadratic or parallel forms. We develop an asymptotic theory for the test statistic. A Gaussian multiplier testing procedure is proposed and it has an improved finite sample performance. Our testing procedure is applied to a spatial temporal temperature data gathered from various locations across America. A simulation study is also presented to illustrate the performance of our testing method.

E0526: **Estimation of the change in variance for Gaussian stationary sequences**
Presenter: **Farida Enikeeva**, University of Poitiers, Laboratoire de Mathematiques et Applications, France
Co-authors: Clement Chesseboeuf , Hermine Bierme
The problem of estimation of a change in variance in a stationary Gaussian sequence is considered. A class of contrast functions based on power variations of the underlying process is considered. We establish a functional limit theorem and then study the properties of the change-point estimator based on the power variations. We will also show how to apply these results to the case of a change in the Hurst parameter of a fractional Brownian motion.

E0923: **Regression with correlated noise: Non-parametric approach**
Presenter: **Paulo Serra**, Eindhoven University of Technology, Netherlands
Co-authors: Tatyana Krivobokova, Francisco Rosales Marticorena
Regression models, particularly of the "signal+noise" variant, play a central role in statistics and are a fundamental tool in many applied fields. Typically, the noise terms are assumed to be independent but this is often not realistic. Methods for selecting bandwidths/smoothing parameters for kernel/spline estimators can break down even if the correlation is mild. Two common approaches are to either "robustify" the criteria for selecting bandwidth/smoothing parameters, or making a parametric assumption on the noise. Unfortunately, both approaches are sensitive to misspecification. We will focus on a non-parametric approach using smoothing spline estimators. The spline parameters and correlation matrix are estimated via the empirical Bayes approach. The estimation of the correlation is rather non-trivial due to the unknown mean of the data. We will consider some implementation issues, and the asymptotics of the estimators. These asymptotics make explicit the influence of the correlation

structure on the smoothing parameters of the penalised spline, and introduce some non-trivial constraints on the order of the splines. We will close with some numerical experiments where we compare our approach to competing estimators, and to a standard R procedure based on a (parametric) assumption on the noise structure.

E1378: **Nonparametric estimation in spatial regression**
Presenter: **Tatiyana Apanasovich**, George Washington University, United States
It is well known that the selection of the smoothing parameter in nonparametric regression is difficult when the errors are spatially correlated. We will discuss various smoothing parameter selection procedures which require a prior knowledge about the correlation structure. Next, we propose a regression estimation framework based on modified kernels which is less sensitive to correlated errors and requires little to no prior knowledge about their correlation structure and its parameters. We demonstrate the practical value of the proposed methodology through simulation studies and real data analysis.

E0384: **Kernel partial least squares for stationary data**
Presenter: **Marco Singer**, Georg-August-Universitaet Goettingen, Germany
Co-authors: Tatyana Krivobokova, Axel Munk
The kernel partial least squares algorithm for nonparametric regression with stationary dependent data is considered. Probabilistic convergence rates of the kernel partial least squares estimator to the true regression function are established when the algorithm is stopped early. The convergence rates depend on three quantities: the regularity of the target function given by a source condition, the effective dimensionality of the data mapped into the reproducing kernel Hilbert space and the the range of dependence in the data measured via the polynomial decay of the autocovariance function. It is shown both theoretically and in simulations that long range dependence results in slower convergence rates.

EO162 Room G3 STATISTICS FOR HIGH FREQUENCY DATA: THEORY AND APPLICATIONS Chair: Nakahiro Yoshida

E0546: **Modeling time scale in high-frequency data**
Presenter: **Hiroki Masuda**, Kyushu University, Japan
Co-authors: Shoichi Eguchi
The aim focuses on when and how time scale can be quantified based on high-frequency sample from a diffusion process. On the one hand, it is well-known that under the standard regularity conditions on the coefficients, the Gaussian quasi-likelihood efficiently works. On the other hand, however, one would get confused with the practical problem what value is to be assigned to the sampling stepsize, for there is no absolute correspondence between model-time and actual-time scales. In this respect, it should be noted that the sampling stepsize is a fine-tuning parameter affecting estimates. Although we could select it in a subjective manner to meet the theoretical requirements on the sampling frequency, it is obviously convenient to have a way to assign a specific value to the stepsize in a data-adaptive manner. We will propose a modified Gaussian quasi-likelihood function which is completely free from the stepsize fine-tuning, and also leads to the following desirable properties under an additional, seemingly non-standard identifiability condition. (1) The associated estimator is rate-efficient and asymptotically normally distributed. (2) The sampling stepsize can be estimated in some sense.

E0556: **Hybrid estimators for ergodic diffusion processes based on thinned data**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Hybrid estimation is considered for both drift and diffusion coefficient parameters for discretely observed ergodic diffusion processes. In order to get the maximum likelihood type estimator, it is crucial to choose a suitable initial estimator for optimization of the quasi likelihood function. From a computational point of view, an initial Bayes type estimator of the diffusion coefficient parameter is given by using reduced data obtained from full data, and an initial Bayes type estimator of the drift parameter is obtained by using thinned data out of full data. The adaptive maximum likelihood type estimator with the initial Bayes type estimator, which is called hybrid estimator, is proposed. The asymptotic properties of the initial Bayes type estimators are proved. Moreover, it is shown that the hybrid estimators based on the initial Bayes type estimators have asymptotic normality and convergence of moments. In order to investigate asymptotic performance of the proposed estimators, we give a concrete example and simulation results.

E1136: **Lead-lag analysis of non-synchronously observed time series with R**
Presenter: **Yuta Koike**, Tokyo Metropolitan University, Japan
A lead-lag relationship is a time-lagged correlation structure of two time series wherein one is correlated to the other with a delay. It has also been well-recognized since long ago that the non-synchronicity of observation times of two time series causes a serious bias in estimating lead-lag relationships. To overcome this issue, a recent work has introduced a novel approach to compute the cross-covariances of the returns of two non-synchronously observed time series as well as proposed a simple statistic to measure their lead-lag relationship. The methodology is not only applicable to high-frequency financial data but also applicable to SNS data, so it could provide a useful tool for lead-lag analysis of time series to empirical researchers in any areas. R package yuima provides systematic functions to conveniently apply this methodology to real time series. The aim is to present them to empirical researchers as well as to show what we can really do in yuima. As an illustration, we will demonstrate its application to real data.

E1351: **The Yuima framework for simulation and inference of stochastic processes and its GUI**
Presenter: **Emanuele Guidotti**, University of Milan, Italy
Co-authors: Stefano Iacus
The purpose is to present the Yuima package, a system of S4 classes and methods for the simulation and inference of stochastic processes including stochastic differential equation with or without jumps, fractional Brownian motion, Poisson and general point processes, CARMA and COGARCH processes. Yuima is a collaborative project and includes several simulation schemes as well statistical tools for quasi-maximum likelihood estimation, model selection, hypotheses testing, change point analysis. It also include methods of asymptotic expansion. Recently, the Yuima package has been coupled with a graphical user interface, namely the YuimaGUI, which simplifies the usage of the package and allows for a complete flow of analysis: from data ingestion, to model selection and or estimation, and estimation.

EO031 Room G4 NEW METHODS FOR ANALYZING COMPLEX DATA Chair: Yue Niu

E0352: **Gaussianity test for high-dimensional data**
Presenter: **Hao Chen**, University of California at Davis, United States
Co-authors: Yin Xia
Many high-dimensional data analysis tools require the data have Gaussian or sub-Gaussian tails. We here provide a general framework for testing whether the data have Gaussian tail or heavier tail. The method extends from graph-based two-sample tests and work when a reasonable covariance matrix estimation is possible. Under some mild conditions on the covariance matrix estimation, the test is consistent against all distributions with tail heavier than the Gaussian distribution.

E0345: **Weighted adaptive hard threshold signal approximation for robust change point detection**
Presenter: **Xiaoli Gao**, University of North Carolina at Greensboro, United States

The copy number detection is incorporated into a change point regression model and a new robust change point detection method is proposed to simultaneously identify breakpoints and outliers. This new model incorporates an individual weight for each observation and uses the adaptive hard threshold approach to efficiently locate both outliers and CNVs. A novel way to select tuning parameters is also adopted in this model. The performance of the proposed change point method is demonstrated by both simulations and real data analysis. We show that the new model preforms more accurately in detecting the number of true break points, particularly given noisy data sets. As a by-product, the proposed approach can detect outliers efficiently.

E0500: **Adaptive concentration and consistency of tree-based survival models**
Presenter: **Yifan Cui**, University of North Carolina at Chapel Hill, United States
Co-authors: Ruoqing Zhu, Mai Zhou, Michael Kosorok

As one of the most popular extensions of random forests, tree-based survival models lack established theoretical results and unified theoretical framework. We investigate the method from the perspective of splitting rules, where the log-rank test statistics is calculated and compared. The splitting rule is essentially treating both resulting child nodes as being identically distributed. However, we demonstrate that this approach is affected by censoring, which may lead to inconsistency of the method. Based on this observation, we develop the adaptive concentration bound result for tree and forest versions of the survival tree models, and establish a general framework for showing the consistency of tree-based survival models. Interestingly, we also show that existing methods based on this biased selection of splitting rule can still lead to consistency as long as the censoring effect is weak.

E1794: **Reduced ranked linear discriminant analysis**
Presenter: **Yue Niu**, University of Arizona, United States
Co-authors: Ning Hao, Bin Dong

Many high dimensional classification techniques have been developed recently. However, most works focus on only the binary classification problem. Available classification tools for the multi-class cases are either based on over-simplified covariance structure or computationally complicated. Following the idea of reduced ranked linear discriminant analysis, we introduce a new dimension reduction tool with the flavor of supervised principal component analysis. The proposed method is computationally efficient and can incorporate the correlation structure among the features. We illustrate our methods by simulated and real data examples.

EO296 Room G5 PREDICTIVE MODELS FOR ACTUARIAL SCIENCE Chair: Katrien Antonio

E1602: **A data driven strategy for the construction of tariff classes in P&C insurance**
Presenter: **Roel Henckaerts**, KU Leuven, Belgium
Co-authors: Katrien Antonio, Roel Verbelen, Maxime Clijsters

A fully data driven strategy to incorporate continuous risk factors and geographical information in an insurance tariff is presented. A framework is developed that aligns flexibility with the practical requirements of an insurance company. Our strategy is illustrated with an example from property and casualty (P&C) insurance, namely a motor insurance case study. We start by fitting generalized additive models (GAMs) to the number of reported claims and their corresponding severity. These models allow for flexible statistical modeling in the presence of different types of risk factors: categorical, continuous and spatial risk factors. The goal is to bin the continuous and spatial risk factors such that categorical risk factors result which capture the effect of the covariate on the response in an accurate way, while being easy to use in a generalized linear model (GLM). This is in line with the requirement of an insurance company to construct a practical and interpretable tariff. We propose to bin the spatial risk factor using Fishers natural breaks algorithm and the continuous risk factors using evolutionary trees. GLMs are fitted to the claims data with the resulting categorical risk factors. We find that the resulting GLMs approximate the original GAMs closely, and lead to a very similar premium structure.

E1771: **Estimating the maximum possible earthquake magnitude using extreme value methodology: The Groningen case**
Presenter: **Tom Reynkens**, KU Leuven, Belgium
Co-authors: Jan Beirlant, Andrzej Kijko, John Einmahl

The area-characteristic, maximum possible earthquake magnitude is required by the earthquake engineering community, disaster management agencies and the insurance industry. The Gutenberg-Richter law predicts that earthquake magnitudes follow a truncated exponential distribution. In the geophysical literature several parametric and non-parametric estimation procedures were proposed. Estimation of the maximum possible earthquake magnitude is of course an extreme value problem to which the classical methods for endpoint estimation could be applied. We argue that recent methods on truncated tails at high levels constitute a more appropriate setting for this estimation problem. We present upper confidence bounds to quantify the uncertainty of the point estimates. We also compare methods from the extreme value and geophysical literature through simulations. Finally, the different methods are applied to the magnitude data for the earthquakes induced by gas extraction in the Groningen province of the Netherlands.

E1614: **A time change strategy to model reporting delay dynamics inclaims reserving**
Presenter: **Jonas Crevecoeur**, KU Leuven, Belgium
Co-authors: Katrien Antonio, Roel Verbelen

Holding sufficient capital is essential for an insurance company to ensure its solvability. Hence, predicting the amount of capital needed to fulfil liabilities with respect to past exposure years in a stable and accurate way, is an important actuarial task. Recent research puts focus on the use of detailed information regarding the development of individual claims. This is called the micro-level or granular reserving approach. Reserve calculations are required for both claims that are not yet settled as well as for claims that have already occurred in past exposure years, but which have not yet been reported to the insurer. We focus on estimating the count of these latter claims, by modeling the time between occurrence and reporting of claims, the so-called reporting delay. Using data at daily level we propose a micro-level model for the heterogeneity in reporting delay caused by calendar day effects, such as the weekday pattern and holidays. Previous work on delay features is extended. These methods are illustrated by case studies using multiple real life insurance datasets.

E1770: **Sparse modeling of risk factors in insurance analytics**
Presenter: **Sander Devriendt**, KU Leuven, Belgium
Co-authors: Katrien Antonio, Roel Verbelen, Edward Frees

Insurance companies use predictive models for a variety of analytic tasks, including pricing, marketing campaigns, claims handling, fraud detection and reserving. Typically, these predictive models use a selection of continuous, ordinal, nominal and spatial predictors to differentiate risks. Such models have to be competitive, interpretable by stakeholders and easy to implement and maintain in a production environment. That is why current actuarial literature puts focus on GLMs where risk cells are constructed by binning predictors up front, using ad hoc techniques or professional expertise. Penalized regression is often used to encourage the selection and fusion of predictors in predictive modeling but most penalization

strategies work only when all predictors are of the same type, such as LASSO for continuous variables and Fused LASSO for ordered variables. We design an estimation strategy for GLMs which includes variable selection and the binning of predictors through L1-type penalties. We consider the joint presence of different types of predictors with their respective penalties. Using the theory of proximal operators, our estimation procedure is computationally efficient since it splits the overall optimization problem into easier sub-problems per predictor and its penalty. We illustrate through simulations and a motor-insurance case-study that we are able to build a sparse regression model, in a statistically sound way, for data with different types of predictors.

EO471 Room Gordon BIG DATA: CONVERGENCE OF STATISTICS AND OPTIMIZATION Chair: Stephane Chretien

E1516: **Machine learning with universal reservoir computers using non-homogeneous state-affine systems and forecasting tasks**
Presenter: **Lyudmila Grigoryeva**, University of Konstanz, Germany
Co-authors: Juan-Pablo Ortega

A new class of non-homogeneous state-affine systems is proposed which can be used for various machine learning applications. Sufficient conditions are identified that guarantee first, that the associated reservoir computers with linear readouts are causal, time-invariant, and satisfy the fading memory property and second, that a subset of this class is universal in the category of fading memory filters with stochastic bounded inputs. This means that any discrete-time filter that satisfies the fading memory property with random inputs of that type can be uniformly approximated by elements in the non-homogeneous state-affine family. We empirically demonstrate the competitive performance of the proposed non-homogeneous state-affine systems in time series forecasting tasks.

E1560: **From safe screening rules to working sets for faster Lasso-type solvers**
Presenter: **Mathurin Massias**, INRIA Saclay, France
Co-authors: Alexandre Gramfort, Joseph Salmon

Convex sparsity-promoting regularizations are ubiquitous in modern statistical learning. By construction, they yield solutions with few non-zero coefficients, which correspond to saturated constraints in the dual optimization formulation. Working set (WS) strategies are generic optimization techniques that consist in solving simpler problems that only consider a subset of constraints, whose indices form the WS. Working set methods therefore involve two nested iterations: the outer loop corresponds to the definition of the WS and the inner loop calls a solver for the subproblems. For the Lasso estimator a WS is a set of features, while for a Group Lasso it refers to a set of groups. In practice, WS are generally small in this context so the associated feature Gram matrix can fit in memory. Here we show that the Gauss-Southwell rule (a greedy strategy for block coordinate descent techniques) leads to fast solvers in this case. Combined with a working set strategy based on an aggressive use of so-called Gap Safe screening rules,we propose a solver achieving state-of-the-art performance on sparse learning problems. Results are presented on Lasso and multi-task Lasso estimators.

E1830: **New tests for generalized linear models based on lasso**
Presenter: **Sylvain Sardy**, University of Geneva, Switzerland

We propose new statistics based on glm-lasso to test a null-hypothesis. We show the statistics is asymptotically pivotal under H_0. The new test behaves better than the classical deviance test (better level and power) when $P < N$ and can still be applied when $P \geq N$ (more covariates then observations).

E1842: **Multiple change-point estimation in Gaussian graphical models**
Presenter: **Sandipan Roy**, University College London, United Kingdom
Co-authors: Alex Gibberd

Consistency properties are considered for a regularised estimator for the simultaneous identification of both change points and dependency structure in multivariate time-series. Traditionally, the estimation of Gaussian Graphical Models (GGM) is performed in an i.i.d setting. More recently such models have been extended to allow for changes in the distribution, but only where change points are known a-priori. We study the Group-Fused Graphical Lasso (GFGL), which penalises partial-correlations with an l1 penalty, while simultaneously inducing block-wise smoothness over time to detect multiple change points. We present a proof of change-point consistency for the estimator. In particular, we demonstrate that both the change-point and graphical structure of the process can be consistently recovered and provide finite sample bounds.

EO674 Room CLO 101 J-ISBA SESSION: NONPARAMETRIC BAYESIAN ANALYSIS OF COPULA MODELS Chair: Giovanna Jona Lasinio

E0256: **Bayesian nonparametric conditional copula estimation of twin data**
Presenter: **Luca Rossini**, Free University of Bozen, Italy
Co-authors: Fabrizio Leisen, Luciana Dalla Valle

Several studies on heritability in twins aim at understanding the different contribution of environmental and genetic factors to specific traits. Considering the National Merit Twin Study, the purpose is to correctly analyse the influence of the socioeconomic status on the relationship between twins' cognitive abilities. The methodology is based on conditional copulas, which allow us to model the effect of a covariate driving the strength of dependence between the main variables. We propose a flexible Bayesian nonparametric approach for the estimation of conditional copulas, which can model any conditional copula density. The methodology extends previous work by introducing dependence from a covariate in an infinite mixture model. The results suggest that environmental factors are more influential in families with lower socio-economic position.

E0438: **A nonparametric Bayesian approach to copula estimation**
Presenter: **Shaoyang Shaoyang Ning**, Harvard University, United States
Co-authors: Neil Shephard

A novel Dirichlet-based Polya tree (D-P tree) prior on the copula is proposed, and based on the D-P tree prior, a nonparametric Bayesian inference procedure is developed. Through theoretical analysis and simulations, we are able to show that the flexibility of the D-P tree prior ensures its consistency in copula estimation, thus able to detect more subtle and complex copula structures than earlier nonparametric Bayesian models, such as a Gaussian copula mixture. Further, the continuity of the imposed D-P tree prior leads to a more favorable smoothing effect in copula estimation over classic frequentist methods, especially with small sets of observations. We also apply our method to the copula prediction between the S&P 500 index and the IBM stock prices during the 2007-08 financial crisis, finding that D-P tree-based methods enjoy strong robustness and flexibility over classic methods under such irregular market behaviors.

E0766: **Introducing dependence among Dirichlet processes: A copula based approach**
Presenter: **Gianluca Mastrantonio**, Politecnico of Turin, Italy
Co-authors: Clara Grazian, Enrico Bibbona

The Dirichlet process is a stochastic process defined on a space of distribution functions and which depends on a scaling parameter and a base distribution. A draw form a Dirichlet process is a random distribution which is almost surely discrete. The Dirichlet process may be extended in a hierarchical version, so that several processes share the same set of atoms with process-dependent weights. As the Dirichlet process, its hierarchical

extension has an explicit representation, called stick-breaking, that generates the weights from independent Beta random variables. We propose a way to introduce dependence in the marginal distributions of the vectors of weights, by imposing a copula model with given dependence (for instance, implying spatial or temporal dependence). We also prove that the dependence structure imposed on the (independent) components of the stick-breaking representation is automatically transferred to the vectors of weights and that the order in which the components are taken does not matter.This representation may be used to produce a structured nonparametric clustering, where the weights of the Dirichlet process represents the probabilities to be allocated to each cluster and the spatial or temporal dependence is among the weights relative to the same cluster.

E1103: **Levy copulas in Bayesian non-parametric models**
Presenter: **Alan Riva Palacio**, University of Kent, Mexico
Co-authors: Fabrizio Leisen
Completely random measures have been used to construct a variety of models in Bayesian non-parametric statistics. Generalization of such models into a multivariate setting can be done in the framework of Levy copulas. We highlight the family of Levy copulas defined by certain compound random measures. In particular we showcase a generalization of the Clayton Levy copula. We present an application to vectors of survival functions where inference is done in a fully Bayesian way that takes advantage of simulation algorithms associated to Levy copulas.

EO166 Room CLO 102 RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS Chair: Anne Ruiz-Gazen

E0816: **ICS with singular scatter matrices**
Presenter: **Aurore Archimbaud**, Toulouse School of Economics, France
Co-authors: Anne Ruiz-Gazen, Klaus Nordhausen
The Invariant Coordinate Selection method is aimed at detecting some potential structure of groups or the presence of outliers in a multivariate data base and in an unsupervised way. ICS is based on the joint diagonalisation of two scatter matrices which are assumed to be non-singular. In practice, however, because of some multicollinearity problems in high dimension, the scatter matrices may be singular. In such a context, it is possible to generalize ICS by using some Generalized Singular Value Decomposition. This approach has some advantages in particular compared with another approach based on generalized inverse of scatter matrices. In some examples where the structure of interest is contained in some subspace, the proposed method is able to recover the subspace of interest while other approaches may fail in identifying such a subspace. These advantages are discussed in detail from a theoretical point of view and using some simulated examples.

E0925: **Application of BigVAR: Forecasting fund shares using social variables**
Presenter: **Viktoria Oellerer**, PwC Vienna, Austria
Issuers of funds need to forecast the purchase and disposal of fund shares of their clients. This is especially important for real estate funds, where selling/buying underlying assets takes a considerable amount of time. As a solution, one can use traditional time-series models relying on macroeconomic variables (e.g. vector autoregression models). In todays big data era, it has become quite common to use a large number of social factors trying to model the behavior of people. We want to combine these two ideas by forecasting the purchase and disposal of fund shares using both macroeconomic and social variables. The presented approach builds on the VARX-L, a regularized vector autoregression model allowing for exogenous variables. The regularization aspect of VARX-L improves forecasting accuracy, the introduction of exogenous variables facilitates stress testing.

E0976: **Asymptotic comparison of tensorial and vectorial ICA methods**
Presenter: **Joni Virta**, University of Turku, Finland
Co-authors: Bing Li, Klaus Nordhausen, Hannu Oja
Naturally tensor-valued data is more and more common nowadays and various different probabilistic tensorial models have been developed in response. The tensor independent component model and various methods for estimating its parameters are discussed. The methods can be divided into two classes, tensorial and those requiring vectorization of the data prior to use, a procedure that causes loss of structural information. To see how severe the loss is the different methods are compared based on their asymptotic efficiencies, yielding the conclusion that under the model vectorization is practically always inferior to the purely tensorial ICA methods.

E1230: **Testing the dimension of the non-Gaussian subspace in NGCA**
Presenter: **Klaus Nordhausen**, Vienna University of Technology, Austria
Co-authors: Hannu Oja, David Tyler, Joni Virta
Dimension reduction is often a preliminary step in the analysis of large data sets. The so-called non-Gaussian component analysis (NGCA) searches for a projection onto the non-Gaussian part of the data, and it is then important to know the correct dimension of the non-Gaussian signal subspace. Different tests for the dimension based on the fourth order blind identification (FOBI) method are presented.

EO407 Room Jessel NEW TRENDS IN ROBUST ESTIMATION Chair: Adrien Saumard

E0457: **Learning from MOM's principle: Theoretical results**
Presenter: **Matthieu Lerasle**, CNRS, France
Co-authors: Guillaume Lecue
A new estimator obtained by minimaximization of Median-Of-Means (MOM) criteria is presented which can be regularized when some structure is assumed on the signal to be estimated. This estimator is particularly relevant in an environment corrupted by outliers, that may be as aggressive as they want. Informative data (i.e. data that are not outliers) are only asked to be independent (not necessarily identically distributed) and to satisfy a weak L2 / L1 moment comparison assumption. In this setting, our estimator performs as well as the ERM and its regularized version would do in a friendly i.i.d. sub-Gaussian environment provided that the number of outliers does not exceed (number of data)*(minimax rate of convergence). These performances are achieved for many risk functions, even for some quite sensitive to the presence of outliers as the quadratic loss, as well as for any regularization norm. A particular emphasis will be put on the l_1, S_1 and SLOPE norms.

E0454: **Learning from MOM's principle**
Presenter: **Guillaume Lecue**, CNRS and ENSAE, France
Co-authors: Matthieu Lerasle
This is the second part of a session on a mini-maximization Median of Means estimator. In this second part, we present several algorithms to compute our estimator and discuss practical performances in various settings. Several points of interest will be stressed: 1) many algorithms designed for the ERM and its regularized versions can easily be adapted to compute the MOMs versions of these estimators (we will present in details an alternating ADMM algorithm designed for the MOM estimator); 2) these MOMs versions inherit robustness properties of MOMs estimators, in particular they are naturally resistant to outliers, heavy-tailed and non identically distributed data; 3) when applied to computationally heavy non-linear methods, MOMs version, operating on smaller batch of data, are much faster to implement. For some large distributed data-sets, the MOM estimator is tractable on a personal computer while the original version requires much stronger computational powers; 4) compared to

other robust estimators, MOMs estimators can naturally be fixed to avoid local maxima; 5) we also derive a concept of outlier importance for the problem of outliers detection is naturally associated to a notion of centrality of data. Illustrations of those points will be outlined in particular by a MOM version of the classical LASSO.

E0863: **On hazard rate estimation for censored data**
Presenter: **Mathieu Sart**, Univ Lyon, UJM-Saint-Etienne, Institut Camille Jordan, France
The hazard rate is proposed to be estimated from censored data by broadening a recent estimation method based on models and named rho-estimation. As we shall see, this method may reduce to the one of maximum likelihood under suitable assumptions. However, the two methods differ in general. In particular, rho-estimation also applies to numerous models where the maximum likelihood method does not work. Besides, rho-estimators generally possess better robustness properties. As a hazard rate cannot be uniformly estimated on the positive real line, we propose to evaluate the quality of the estimators by means of a random Hellinger-type loss adapted to the statistical setting. We establish non-asymptotic risk bounds that are very similar to the ones that can be obtained in density estimation. We then specify these bounds under some assumptions on the hazard rate.

E1408: **Concentration for robust mean estimators: Some recent results**
Presenter: **Emilien Joly**, CIMAT, Mexico
The estimation of the mean is a crucial step in many statistical contexts. Huber's work on the study of the asymptotic behavior of robust estimators in the context of heavy tailed random variables has recently been revisited. New robust estimators have been proposed that satisfy concentration inequalities hence at finite and fixed n with sub-Gaussian speed under a small moment assumption (finite variance). The robust estimation of the mean for real valued random variables is a starting point for extensions of robust estimators to more complex probability fields. The focus will be on the definition of a robust estimator and its concentration in the case of vector valued distributions.

EO017 Room Montague RISK PREDICTION IN SURVIVAL ANALYSIS Chair: Ingrid Van Keilegom

E0221: **Smooth time-dependent ROC curve estimators**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Pablo Martinez-Camblor
The ROC curve is a popular graphical tool often used to study the diagnostic capacity of continuous (bio)markers. When the considered outcome is a time-dependent variable, two main extensions have been proposed: the cumulative/dynamic ROC curve and the incident/dynamic ROC curve. In both cases, the main problem for developing appropriate estimators is the estimation of the joint distribution of the variables time-to-event and marker. As usual, different approximations lead to different estimators. We explore the use of a bivariate kernel density estimator which accounts for censored observations and produces smooth estimators of the time-dependent ROC curves. The performance of the resulting cumulative/dynamic and incident/dynamic ROC curves is studied by means of Monte Carlo simulations. Additionally, the influence of the choice of the required smoothing parameters is explored. The proposed methodology is illustrated with applications to real data.

E1060: **Evaluating the accuracy of prognostic biomarkers in the presence of external information**
Presenter: **Maria Xose Rodriguez-Alvarez**, Universidade de Vigo, Spain
Co-authors: Thomas Kneib
Prior to using a diagnostic biomarker in a routine clinical setting, the rigorous evaluation of its diagnostic accuracy is essential. The Receiver Operating Characteristic (ROC) curve is the measure of accuracy most widely used for continuous biomarkers. However, the possible impact of extra information about the patient (or even the environment) on diagnostic accuracy also needs to be assessed. In addition, in many circumstances the aim of a study may involve prognosis rather than diagnosis. The main difference between diagnostic and prognostic biomarkers is that, with prognostic biomarkers, a time dimension is involved. This is the case of survival studies, where the status of an individual varies with time (e.g, death and alive). To assess the accuracy of continuous prognostic biomarkers for time-dependent disease outcomes, time-dependent extensions of the ROC curve have been proposed. A novel penalized likelihood-based estimator of the cumulative-dynamic time-dependent ROC curve is presented. The proposal allows to account for the possible modifying effect of covariates on the accuracy of the biomarker. The validity of the approach is supported by simulations, and applied to the evaluation of biomarkers for early prognosis of death after discharge in patients who suffered an acute coronary syndrome.

E1289: **Simultaneous modeling of counts with excess zeros and left-truncated survival data with time-varying effects**
Presenter: **Xavier Piulachs**, Universite Catholique de Louvain - University of Barcelona, Belgium
Co-authors: Montserrat Guillen
In the context of health care studies, repeated measurements of medical claims are collected on each subject as an indirect measure of health status. At the same time, the aging process taking place in developed countries leads to an obvious interest in assessing the relationship between emergency care demand and survival rate, paying particular attention to elderly policyholders. We propose a joint model to analyze the degree of dependence between the longitudinal and time-to-death outcomes for a large cohort of Spanish policyholders aged 65 and over. The longitudinal response is restricted to a small range of non-negative integer values, and is affected by some degree of overdispersion caused by unobserved heterogeneity and an excess of zeros. Additionally, only policyholders who have reached the age of 65 come under study, so death times may be subject to both left truncation and right censoring. Finally, we argue the existence of a time-varying relationship between longitudinal and survival responses. The estimation of parameters is performed under a Bayesian perspective by Markov chain Monte Carlo simulation methods.

E0261: **The balance hazard ratio for evaluating prognostic factors and survival risk groups**
Presenter: **Pierre Dupont**, University of Louvain, Belgium, Belgium
Common clinical studies assess the quality of prognostic factors, such as gene expression signatures, clinical variables or environmental factors to cluster patients into various risk groups. Such groups are intended to represent patients with similar survival odds and to select the most appropriate therapy accordingly. The relevance of such risk groups, and of the related prognostic factors, is typically assessed through the computation of a hazard ratio. Some limitations of this commonly adopted procedure are stressed here. These issues could lead to inappropriate comparisons between various prognostic factors. One briefly describes some alternative measures including the Concordance Index, AUC, ... and discuss their benefits and limitations. Next, the balanced hazard ratio is introduced to solve these issues. This new performance metric keeps an intuitive interpretation and is as simple to compute as the original hazard ratio. It also leads to a natural cut-off choice to define risk groups from continuous risk scores. The proposed methodology is illustrated through the evaluation of a gene signature for breast cancer prognosis, and similar results are briefly presented on alternative clinical studies. Finally, one argues why the proposed methodology can be applied more generally to assess the quality of any prognostic or predictive markers.

EO216 Room Senate OUTLIERS AND CHANGE-POINTS IN TIME SERIES Chair: Roland Fried

E0677: **Testing normality of functional time series**
Presenter: **Siegfried Hoermann**, Univ libre de Bruxelles, Belgium
Co-authors: Piotr Kokoszka, Lajos Horvath, Tomasz Gorecki
Tests of normality for a time series of functions are developed. The tests are related to the commonly used Jarque Bera test. The assumption of normality has played an important role in many methodological and theoretical developments in the field of functional data analysis, yet, no inferential procedures to verify it have been proposed so far, even for iid functions. We propose different approaches and evaluate the tests via simulations. For the test which gives the best size and power trade-off we establish large sample validity under general conditions. We obtain interesting insights by applying them to pollution and intraday price curves. While the pollution curves can be treated as normal, the normality of high frequency price curves is rejected.

E0699: **Multiple change point estimation based on moving sum statistics**
Presenter: **Claudia Kirch**, Otto-von-Guericke University Magdeburg, Germany
Using the example of changes in the mean we introduce change point estimators for multiple changes based on moving sum statistics. For a given bandwidth we prove consistency of these estimators as well as their asymptotic limit distribution. In many practical situations a single bandwidth is not sufficient to detect all change points, but instead several bandwidths need to be used. We discuss extensions based on information criteria to combine this multiscale information in a consistent way. Finally, extensions to a very general setting based on estimating functions are discussed. These extensions allow for more robust methods in the mean change situation as well as for more complex change point problems such as changes in the time series structure.

E0883: **Detecting and dating structural breaks in functional data without dimension reduction**
Presenter: **Ozan Sonmez**, University of California, Davis, United States
Co-authors: Gregory Rice, Alexander Aue
Methodology is proposed to uncover structural breaks in functional data that is fully functional in the sense that it does not rely on dimension reduction techniques. A thorough asymptotic theory is developed for a fully functional break detection procedure as well as for a break date estimator, assuming a fixed break size and a shrinking break size. The latter result is utilized to derive confidence intervals for the unknown break date. The main results highlight that the fully functional procedures perform best under conditions when analogous fPCA based estimators are at their worst, namely when the feature of interest is orthogonal to the leading principal components of the data. The theoretical findings are confirmed by means of a Monte Carlo simulation study in finite samples. An application to annual temperature curves illustrates the practical relevance of the proposed fully functional procedure. An R function is also developed for structural break analysis in functional time series, and its functionalities will be discussed.

E1016: **Studentized U-quantile processes**
Presenter: **Daniel Vogel**, University of Aberdeen, United Kingdom
Co-authors: Carina Gerstenberger, Martin Wendler
Functional limit theorems for U-quantiles of stationary, but potentially dependent series are studied, and a consistent estimator for the long-run variance is proposed. This allows us to formulate studentized versions of the processes, which are asymptotically distribution-free. Such results are useful, e.g., for change-point analysis. The Hodges–Lehmann estimator and the Q_n estimator are popular estimators for location and scale, respectively. Both combine high efficiency at normality with appealing robustness properties, and both are U-quantiles. We propose and investigate change-point test statistics based on these estimators for detecting changes in the central location or the scale of time series. A strength of quantile-based estimators in general is "moment-freeness". We use the notion of near epoch dependence in probability (PNED), a very general weak dependence condition, which, in contrast to the traditional L_2 near epoch dependence, does not require the existence of any moments, making it particularly appealing for the analysis of robust methods under dependence.

EO492 Room Woburn MODERN STATISTICAL METHODS FOR COMPLEX DATA Chair: Yang Feng

E0727: **Testing goodness-of-fit with complex data**
Presenter: **Ming-Yen Cheng**, Hong Kong Baptist University and National Taiwan University, China
In the emerging era of big data, it occurs often that both the dimensionality and complexity of the data are high. Although parametric models enjoy good efficiency and interpretability if they are correctly specified, they suffer from large bias if they are incorrect. Nonparametric models are flexible in allowing the data to speak for the underlying structure and can serve as useful tools for model checking. However, they run into the curse of dimensionality problem and the variance dominates. In addition, in some cases identification is a problem. Semiparametric models offer a trade-off between parametric and nonparametric models and enjoy advantages of both of the two. Therefore they are particularly useful in analysis of complex data. We will focus on goodness-of-fit testing for some semiparametric regression models. Theoretical and numerical results, along with applications to data sets coming from medical and climate studies will be given to demonstrate the efficacy and advantages of the proposed methods.

E1112: **Optimal treatment allocation for emerging infectious diseases**
Presenter: **Eric Laber**, North Carolina State University, United States
A key component in controlling the spread of an epidemic is deciding where, when, and to whom to apply an intervention. We develop a framework for using data to inform these decisions in real-time. We formalize a treatment allocation strategy as a sequence of functions, one per treatment period, that map up-to-date information on the spread of an infectious disease to a subset of locations where treatment should be allocated. An optimal allocation strategy optimizes some cumulative outcome, e.g., the number of uninfected locations, the geographic footprint of the disease, or the cost of the epidemic. Estimation of an optimal allocation strategy for an emerging infectious disease is challenging because spatial proximity induces interference among locations, the number of possible allocations is exponential in the number of locations, and because disease dynamics and intervention effectiveness are unknown at outbreak. We derive a Bayesian online estimator of the optimal allocation strategy that combines simulation-optimization with Thompson sampling.

E0621: **High-dimensional changepoint estimation via sparse projection**
Presenter: **Tengyao Wang**, University of Cambridge, United Kingdom
Co-authors: Richard Samworth
Changepoints are a very common feature of Big Data that arrive in the form of a data stream. We study high-dimensional time series in which, at certain time points, the mean structure changes in a sparse subset of the coordinates. The challenge is to borrow strength across the coordinates in order to detect smaller changes than could be observed in any individual component series. We propose a two-stage procedure called 'inspect' for estimation of the changepoints: first, we argue that a good projection direction can be obtained as the leading left singular vector of the matrix that solves a convex optimisation problem derived from the CUSUM transformation of the time series. We then apply an existing univariate changepoint

estimation algorithm to the projected series. Our theory provides strong guarantees on both the number of estimated changepoints and the rates of convergence of their locations, and our numerical studies validate its highly competitive empirical performance for a wide range of data generating mechanisms. Software implementing the methodology is available in the R package 'InspectChangepoint'.

E1231: **On the consistency of graph-based semi-supervised learning**
Presenter: **Yunpeng Zhao**, George Mason University, United States
Co-authors: Chengan Du
Graph-based semi-supervised learning is one of the most popular methods in machine learning. Some of its theoretical properties such as bounds for the generalisation error and the convergence of the graph Laplacian regulariser have been studied in computer science and statistics literatures. However, a fundamental statistical property, the consistency of the estimator from this method has not been proved. We study the consistency problem under a non-parametric framework. We prove the consistency of graph-based learning in the case that the estimated scores are enforced to be equal to the observed responses for the labeled data. The sample sizes of both labeled and unlabeled data are allowed to grow in this result. When the estimated scores are not required to be equal to the observed responses, a tuning parameter is used to balance the loss function and the graph Laplacian regulariser. We give a counterexample demonstrating that the estimator for this case can be inconsistent. The theoretical findings are supported by numerical studies.

EO023 Room CLO 203 THE STEIN METHOD IN STATISTICS: A NEW QUANTIFICATION OF APPROXIMATIONS Chair: Christophe Ley

E0196: **Exact confidence intervals for the Hurst parameter of a fractional Brownian motion**
Presenter: **Ivan Nourdin**, University of Luxembourg, Luxembourg
The aim is to show how to use concentration inequalities arising when one combines Stein's method and Malliavin calculus, in order to build exact confidence intervals for the Hurst parameter associated with a one-dimensional fractional Brownian motion.

E0379: **Differential Stein operators for multivariate continuous distributions and applications**
Presenter: **Gesine Reinert**, Oxford University, United Kingdom
Co-authors: Yvik Swan
The Stein operator approach has been fruitful for one-dimensional distributions. We generalise it to multivariate continuous distributions. We shall discuss different notions for the Stein kernel. Among the applications we consider is the Wasserstein distance between two continuous probability distributions under the assumption of existence of a Poincare' constant.

E0443: **The rate of convergence of some asymptotically chi-square distributed via Stein's method**
Presenter: **Robert Gaunt**, The University of Manchester, United Kingdom
Co-authors: Gesine Reinert
Stein's method is a powerful technique for bounding the distance between two probability distributions with respect to a probability metric. We see how the classical Stein's method for normal approximation can be extended relatively easily via a general transfer principle to derive approximations for statistics that can be expressed as a function of a multivariate normal random variable. We also note a surprising result regarding the rate of convergence. This approach is used to derive a bound on the rate of convergence of Pearson's statistic to its limiting chi-square distribution. The bound has the correct dependence on the model parameters. We end by noting that our approach can also be applied to many other statistics, including Friedman's statistic and the family of power divergence statistics.

E0508: **Measuring the impact of priors with Stein's Method**
Presenter: **Fatemeh Ghaderinezhad**, Gent university, Belgium
Co-authors: Christophe Ley
Recently a new procedure has been proposed to measure the impact of the choice of the prior in Bayesian statistics. This approach is based on the Stein Method in combination with the Wasserstein distance. We will present new results on this topic, by focusing on one-parameter models.

EO152 Room CLO 204 EMPIRICAL PROCESSES AND APPLICATIONS Chair: Eric Beutner

E0376: **Weak convergence of the least concave majorant of estimators for a concave distribution function**
Presenter: **Brendan Beare**, University of California, San Diego, United States
Co-authors: Zheng Fang
The asymptotic behavior of the least concave majorant of an estimator of a concave distribution function is studied under general conditions. In particular, we allow the true concave distribution function to violate strict concavity, so that the empirical distribution function and its least concave majorant are not asymptotically equivalent. We also allow for serially dependent data and a distribution with potentially unbounded support on the nonnegative half-line. Our results are proved by demonstrating the Hadamard directional differentiability of the least concave majorant operator. While the standard bootstrap fails, we show that the rescaled bootstrap of Dumbgen can deliver valid inference.

E0469: **The central limit theorem for empirical processes in $L^p(\mu)$ with applications**
Presenter: **Javier Carcamo**, Universidad Autonoma de Madrid, Spain
The empirical process associated with a random variable X with distribution function F is shown to converge in distribution to the F-Brownian bridge in $L^p(\mu)$ if and only if it is pregaussian. The result is valid for all $1 \leq p < \infty$ and all σ-finite Borel measure μ. Additionally, we determine several easy to check requirements equivalent to the empirical process being pregaussian. In this way, we determine necessary and sufficient conditions to obtain a Donsker theorem for the empirical process in $L^p(\mu)$. When μ is the Lebesgue measure, the convergence of the empirical process amounts to the membership of X to the Lorentz space $\mathcal{L}^{2/p,1}$. A characterization of the finiteness of the second strong moment of the empirical process in $L^p(\mu)$ is also provided. As an application, we obtain the asymptotic distribution of the plug-in estimators of the L^p-distance between distribution functions. The case $p = 1$ corresponds to the so-called L^1-Wasserstein metric.

E0435: **Some asymptotic results for integrated empirical processes with applications to statistical tests**
Presenter: **Aime Lachal**, INSA de Lyon (University of Lyon) / Institut Camille Jordan, France
Co-authors: Sergio Alvarez-Andrade, Salim Bouzebda
Strong approximations of a class of integrated empirical processes are discussed. For this class, the exact rates of the approximations by a sequence of weighted Brownian bridges and a weighted Kiefer process are presented. The arguments are based in part on the famous Komlos-Major-Tusnady approximation. Applications include in particular the two-sample testing procedures together with the change-point problems.

E0523: **Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift**
Presenter: **Jakob Soehl**, TU Delft, Netherlands
Co-authors: Mathias Trabs
As a starting point we prove a functional central limit theorem for estimators of the invariant measure of a geometrically ergodic Harris-recurrent Markov chain in a multi-scale space. This allows us to construct confidence bands for the invariant density with optimal (up to undersmoothing) L^∞-diameter by using wavelet projection estimators. In addition our setting applies to the drift estimation of diffusions observed discretely with fixed observation distance. We prove a functional central limit theorem for estimators of the drift function and finally construct adaptive confidence bands for the drift by using a completely data-driven estimator.

EO719 Room SH349 RECENT ADVANCES FOR INFERENCE IN LATENT VARIABLE MODELS Chair: Estelle Kuhn

E0607: **Estimating the Fisher information matrix in latent variable models**
Presenter: **Maud Delattre**, AgroParisTech, France
Co-authors: Estelle Kuhn
The Fisher information matrix (FIM) plays a key role in statistics. It is crucial in practice not only for evaluating the precision of parameter estimations, but also for statistical testing or experimental designs. In many latent variable models, the evaluation of the observed FIM is non-elementary due to the unobserved variables. Several methods have then been proposed to estimate the observed FIM. Among the most frequently used approaches are Monte-Carlo methods or iterative algorithms derived from the missing information principle. These methods however require to compute second derivatives of the complete data log-likelihood which has some disadvantages from a computational point of view. We propose an estimator of the observed FIM which only requires to compute the first derivatives of the complete data log-likelihood with respect to the parameters. Indeed, we consider the empirical first order moment estimate of the covariance matrix of the score. We derive a stochastic approximation algorithm for computing this estimator when its expression is not explicit. We study the asymptotic properties of the proposed estimator and of the corresponding stochastic approximation algorithm. Some numerical experiments are performed in mixed-effects models and mixture models to illustrate the theoretical results.

E1348: **Stochastic proximal gradient algorithms for penalized mixed models**
Presenter: **Edouard Ollier**, Ecole Normale Suparieure de Lyon, France
Latent variable models are classical tools to model longitudinal data such as in population pharmacokinetic with non-linear mixed effects models. Selection of such model may rely on the use of penalized maximum likelihood estimator for which penalized version of the SAEM algorithm has been developed. Even if these algorithms perform well in practice, there is a lack of theoretical information concerning their convergence. We will present convergence results in the case of a concave likelihood. Moreover, inspired from proximal gradient algorithms of deterministic optimization and Generalized EM algorithm, we will show that the M step could be reduce to a single application of a proximal gradient operator. These results will be illustrated on simulated data and on a real example.

E0996: **Mixed effects modeling and warping for functional data using B-spline**
Presenter: **Emilie Devijver**, CNRS, France
Co-authors: Gerda Claeskens, Irene Gijbels
In functional data the interest is to find a global mean pattern, but also to capture the individual curve differences in phase and amplitude. This can be done conveniently by building in random effects on two levels: a random effect appearing in the warping functions for accounting individual phase variations; and a random effect to deal with individual amplitude variations. Via an appropriate choice of the warping function and B-spline approximations, estimation in the non-linear mixed effects functional model is feasible, and does not require any prior knowledge on landmarks for the functional data. Sufficient and necessary conditions for identifiability of the flexible model are provided. The finite-sample performance of the proposed estimation procedure is investigated in a simulation study, which includes comparisons with an existing method. The added value of the developed method is further illustrated on the analysis of a real data example.

E0505: **A semiparametric extension of the stochastic block model for longitudinal networks**
Presenter: **Tabea Rebafka**, Pierre et Marie Curie University, France
Co-authors: Catherine Matias, Fanny Villers
To model recurrent interaction events in continuous time, an extension of the stochastic block model is proposed where each individual belongs to a latent group and interactions between two individuals follow a conditional inhomogeneous Poisson process whose intensity is driven by the individuals' latent groups. The model is shown to be identifiable and an estimation procedure is proposed based on a semiparametric variational expectation-maximization algorithm. Two versions of the method are developed, using either a nonparametric histogram approach (with an adaptive choice of the partition size) or kernel intensity estimators. The number of latent groups can be selected by an integrated classification likelihood criterion. Finally, the performance of the new procedure is demonstrated on synthetic experiments and real datasets are analysed to illustrate the utility of the approach.

CI007 Room Beveridge Hall INFERENCE IN SEMIPARAMETRIC GENERALIZED SEPARABLE MODELS Chair: Stefan Sperlich

C0165: **Estimation of additive frontier functions with shape constraints**
Presenter: **Lijian Yang**, Tsinghua University, China
Co-authors: Lan Xue
Production frontier is an important concept in modern economics and has been widely used to measure production efficiency. Existing nonparametric frontier models often only allow one or low-dimensional input variables due to "curse-of-dimensionality". We propose a flexible additive frontier model which quantifies the effects of multiple input variables on the maximum output. In addition, we consider the estimation of the nonparametric frontier functions with shape restrictions. Economic theory often imposes shape constrains on production frontier, such as, monotonicity and concavity. A two-step constrained polynomial spline method is proposed to give smooth estimators that automatically satisfy such shape constrains. The proposed method is not only easy to compute, but also more robust to outliers. In theory, the proposed method also enjoys the optimal mean square rate of convergence. We illustrate the proposed method by both simulation studies and an application to the Norwegian farm data. The numerical studies suggest that the proposed method has superior performance by incorporating shape constraints.

C0164: **Smooth backfitting for errors-in-variables additive models**
Presenter: **Byeong Park**, Seoul National University, Korea, South
A new smooth backfitting method and theory for estimating additive nonparametric regression models when the covariates are contaminated by measurement errors is developed. For this, we devise a new kernel function that suitably deconvolutes the bias due to measurement errors as well as renders a projection interpretation to the resulting estimator in the space of additive functions. The deconvolution property and the projection interpretation are essential for a successful solution of the problem. We prove that the method based on the new kernel weighting scheme achieves the optimal rate of convergence in one-dimensional deconvolution problems when the smoothness of measurement error distribution is less than a

threshold value. We find that the speed of convergence is slower than the univariate rate when the smoothness of measurement error distribution is above the threshold, but it is still much faster than the optimal rate in multivariate deconvolution problems. The theory requires a deliberate analysis of the non-negligible effects of measurement errors being propagated to other additive components through backfitting operation. We present the finite sample performance of the deconvolution smooth backfitting estimators that confirm our theoretical findings.

C0166: **Semiparametric estimation of a sample selection model with binary endogenous regressors**
Presenter: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-authors: Patricia Moreno
An alternative method is provided for estimating the impact of an endogenous treatment effect in the presence of sample selection. The estimation method is based on a two-step approach: in the first step, we use parametric regression for estimating the treatment and selection variables, and in the second one we propose an weighted pairwise differences estimator by using the first ones as control functions accounting for both endogeneity and sample selection but assuming an unknown distribution of the errors in the main equation. The proposed estimator is shown to be consistent and asymptotically normal. An empirical application is presented to demonstrate the usefulness of our method.

CO272 Room MAL 151 CONTRIBUTIONS IN LIQUIDITY Chair: Gaelle Le Fol

C1775: **Variation in funding liquidity and financial stability risks**
Presenter: **Gregory Bauer**, Bank of Canada, Canada
Co-authors: Eleonora Granziera
There is evidence that various measures of market liquidity help forecast business cycles. There is also evidence that market liquidity is related to funding liquidity: the ability of financial intermediaries to raise capital and hold levered positions. We examine how time variation in funding liquidity is related to domestic and global financial cycles. As financial cycles may end in financial crises, we also examine whether variation in funding liquidity can help forecast crises. This is of interest as the quantity of private credit (appropriately normalized) is the most robust variable for forecasting crises. We examine whether time variation in the price of credit may increase forecasting power. We construct several different measures of funding liquidity (e.g., betting against beta strategy returns, idiosyncratic volatility, etc.) in a cross section of 18 advanced economies and use panel techniques to assess predictability.

C1777: **Liquidity taking and stock returns**
Presenter: **Milla Siikanen**, Tampere University of Technology, Finland
Co-authors: Sindhuja Ranganathan, Juho Kanniainen
Most of the modern stock exchanges are organized as centralized limit order books, where investors may choose to take or make liquidity by submitting either market or limit orders. We study the relationship between the imbalance of liquidity taking on buy versus sell sides and past and future returns for different investor categories. From our unique data set, we identify trades of different types of investors, namely financial, non-financial, and non-profit institutions, households and foreign investors, and whether the investor was on the market or the limit order side of the transactions. From this, we calculate the imbalance between liquidity taking on buy versus sell side of the limit order book for each investor group. Preliminary results over all investors indicate that higher past returns lead to more liquidity taking on sell side, whereas high returns are preceded by more liquidity taking on buy side. The results for households, and non-financial and non-profit institutions are mainly similar when compared to the results on all the investors. Interestingly, for financial institutions, higher past returns seem to be related to more liquidity taking on buy side. For the heterogeneous group of foreign investors, we find hardly any statistically significant association between imbalance in liquidity taking and returns.

C1542: **Illiquidity and volatility spillover effects in equity markets during and after a financial crisis: An MEM approach**
Presenter: **Yongdeng Xu**, Cardiff University, United Kingdom
Even though the volatility spillover effects in global equity markets have been documented extensively, the transmission mechanism of illiquidity has not been well investigated. We propose a multiplicative error model (MEM) for the dynamics of illiquidity. We empirically study the illiquidity and volatility spillover effects in eight developed equity markets during and after the recent financial crisis. We find that the equity markets are interdependent, both in terms of volatility and illiquidity. Most markets show an increase in volatility and illiquidity spillover effects during the crisis. Furthermore, we find volatility and illiquidity transmission are highly relevant. Illiquidity is a more important channel than volatility in propagating the shocks in the equity markets. Our results show an overall crucial role of illiquidity of US markets in influencing other equity markets illiquidity and volatility. These findings are of importance for policy makers as well as institutional and private investors.

C1480: **From a quote-driven to an order-driven market: The case of the EuroMTS government bond trading platform**
Presenter: **Hanyu Zhang**, University College Dublin, Ireland
Co-authors: Alfonso Dufour
The liquidity of the EuroMTS market is analized before and after an important rule change, which has allowed every market participant to post limit orders and not just designated market makers. We demonstrate that liquidity measures, such as the bid-ask spread and quantity of government bonds available for trading at the best quotes, improve across maturities and countries. We also consider the execution costs for bond trades with a greater size than the minimum quoted size. In particular, we show that the relative bid-ask spread for trading 10 million bonds decreases with the rule change. The proportion of time when the relative bid-ask spread stays low also increases. Several bond portfolios have been constructed, dividing the whole maturity range into three categories: short-term, medium-term and long-term bonds. The results are robust to the on/off-the-run phenomenon and other influential variables, suggesting that greater competition amongst liquidity providers improves liquidity.

CO496 Room MAL 153 FINANCIAL ECONOMETRICS WITH R Chair: David Ardia

C0185: **Forecasting performance of Markov-switching GARCH models: A large-scale empirical study**
Presenter: **David Ardia**, University of Neuchatel, Switzerland
Co-authors: Keven Bluteau, Kris Boudt, Leopoldo Catania
A large-scale empirical study is carried out to evaluate the forecasting performance of Markov-switching GARCH (MSGARCH) models compared with standard single-regime specifications. We find that the need for a Markov-switching mechanism in GARCH models depends on the underlying asset class on which it is applied. For stock data, we find strong evidence for MSGARCH while this is not the case for stock indices and currencies. Moreover, Markov-switching GARCH models with a conditional (skew) Normal distribution are not able to jointly account for the switch in the parameters as well as for the excess of kurtosis exhibited from the data; hence a Markov-switching GARCH model with (skew) Student-t specification is usually required. Finally, accounting for the parameter uncertainty in predictions, via MCMC, is necessary for stock data.

C0189: **Value-at-risk prediction in R with the GAS package**
Presenter: **Leopoldo Catania**, Aarhus BBS, Denmark
Co-authors: David Ardia, Kris Boudt
Generalized Autoregressive Score (GAS) models have been recently proposed as valuable tools for signal extraction and prediction of time series processes with time-varying parameters. For financial risk managers, GAS models are useful as they take the non-normal shape of the conditional distribution into account in the specification of the volatility process. Moreover, they lead to a completely specified conditional distribution and thus to a straightforward calculation of the one-step ahead predictive Value-at-Risk (VaR). It is shown how the novel GAS package for R can be used for Value-at-Risk (VaR) prediction and provides illustration using the series of log-returns of the Dow Jones Industrial Average constituents. Details and code snippets for prediction, comparison and backtesting with GAS models are presented.

C0203: **Accounting for non-normality and luck in fund peer performance evaluation**
Presenter: **Kris Boudt**, Vrije Universiteit Brussel, Belgium
Co-authors: David Ardia
Delegating the investment decision to a group of carefully chosen fund managers requires to evaluate how well the manager performs compared to his/her peers. The question then is not only whether the manager is individually talented, but also whether he/she is more talented than the others. We discuss the statistical tools in the R package PeerPerformane that provide an answer to this question based on the analysis of historical fund returns. In particular, we stress that such a peer performance evaluation requires to take the non-normality of the fund return series into account, and also needs to correct for the possibly large number of false positives, when comparing the fund's performance with a universe of peer funds. We illustrate this for various performance measures, including the fund's alpha residual performance measure and the modified Sharpe ratio.

CO180 Room MAL 414 RECENT DEVELOPMENTS AND APPLICATIONS IN THE ECONOMETRICS OF NETWORKS Chair: Pierre Siklos

C0406: **Networks of volatility spillovers among stock markets**
Presenter: **Eduard Baumohl**, University of Economics in Bratislava, Slovakia
Co-authors: Stefan Lyocsa, Tomas Vyrost, Evzen Kocenda
In a network analysis of 40 developed, emerging and frontier stock markets during the 2006-2014 period, volatility spillovers during both the global financial crisis and tranquil periods are described and modeled. The resulting market interconnectedness is depicted by fitting a spatial model incorporating several exogenous characteristics. We document the presence of significant temporal proximity effects between markets and somewhat weaker temporal effects with regard to the US equity market - volatility spillovers decrease when markets are characterized by greater temporal proximity. Volatility spillovers also present a high degree of interconnectedness, which is measured by high spatial autocorrelation. This finding is confirmed by spatial regression models showing that indirect effects are much stronger than direct effects; i.e., market-related changes in "neighboring" markets (within a network) affect volatility spillovers more than changes in the given market alone, suggesting that spatial effects simply cannot be ignored when modeling stock market relationships. Our results also link spillovers of escalating magnitude with increasing market size, market liquidity and economic openness.

C1221: **The transmission of sovereign and bank credit risk to the non-financial corporate sector in Europe**
Presenter: **Christian Gross**, University of Muenster, Germany
Co-authors: Martin T Bohl, Pierre Siklos
A high-dimensional network of European CDS spreads is modeled to assess the transmission of credit risk to the non-financial corporate sector in Europe. We build on a previous network connectedness approach, which uses variance decompositions in vector autoregressions (VARs) to characterize the dependence structure in the panel of CDS spreads. We derive several connectedness measures that allow us to quantify the degree of credit risk spillovers to non-financial corporations from sovereigns and banks in the Eurozone. A special emphasis of our analysis is on the network structure of CDS spreads during the Global Financial Crisis and the European Debt Crisis.

C0579: **Signed spillover effects building on historical decompositions**
Presenter: **Pierre Siklos**, Wilfrid Laurier University, Canada
The spillover effects of interconnectedness between financial assets is decomposed into both sources of shocks and whether they amplify or dampen volatility conditions in the target market. We use historical decompositions to rearrange information from a VAR which includes sources, direction and signs of effects building on the unsigned forecast error variance decomposition approach. A spillover index based on historical decompositions has simple asymptotic properties, permitting the derivation of analytical standard errors of the index and its components. We apply the methodology to a panel of CDS spreads of sovereigns and financial institutions for the period 2003-2013 and are able to observe both the direction of spillovers and whether they amplify or dampen volatility in the target market.

CO246 Room MAL 415 VOLATILITY MODELING Chair: Giuseppe Storti

C0957: **Comparing density forecasts in a risk management context**
Presenter: **Cees Diks**, University of Amsterdam, Netherlands
Co-authors: Hao Fang
Multivariate and univariate approaches are compared to asses the accuracy of competing density forecasts of a portfolio return in the downside part of the support. We argue that the common practice to perform multivariate forecast comparisons can be problematic in the context of assessing portfolio risk, since better multivariate forecasts do not necessarily correspond to better aggregate portfolio return forecasts. This is illustrated by examples involving (skew) elliptical distributions and an application to daily returns of a number of US stock prices. Additionally, time-varying test statistics and Value-at-Risk forecasts provide empirical evidence for regime changes over the last decades.

C1312: **A conditional coverage test for forecast combination in Value at Risk prediction**
Presenter: **Malvina Marchese**, Universita' di Genova, Italy
The aim is to investigate the issue of Value at Risk forecasting for high dimensional portfolios from a Value at Risk perspective introducing a variation of a previous model averaging strategy. We consider several MGARCH and SV models with different decay rates and propose a general two step model diversification strategy valid for both classes. We rank the volatility models using several loss functions, robust to the choice of the volatility proxy, and we then construct the forecast combination using the Model Confidence Set approach. We consider several combinations of weights and implement a previous CGMM estimator for their estimation. Finally we establish the asymptotic and finite sample distribution of a conditional coverage test that adapts Christoffersen's conditional coverage test to our model averaging framework. The predictive Value at Risk forecasting performance of the models and of the candidate model-combinations is evaluated for a portfolio of 50 stocks traded on the NYSE over the period 2004-2016.

C1382: **Estimation of realized betas in a multi-factor model in presence of noise and asynchronisity**
Presenter: **Sebastien Laurent**, AMU, France
Co-authors: Orimar Sauri, Silvana Acosta
An estimator of realized betas in a multi-factor model is proposed which is robust to microstructure-noise and asynchronisity. The estimator relies on the Cholesky factorization of the realized covariance matrix and, therefore, it also provides positive semidefinite estimates of the covariance matrix. We prove consistency and asymptotic normality of the estimator of the realized betas. Monte Carlo simulations confirm good finite sample properties. We apply our estimator on 52 stocks US stocks and 10 ETFs.

C1387: **Heterogeneous component MEM models for forecasting trading volumes**
Presenter: **Giuseppe Storti**, University of Salerno, Italy
Co-authors: Antonio Naimoli
A novel approach is proposed for modelling and forecasting high-frequency trading volumes, revisiting the Component Multiplicative Error Model by a more flexible specification of the long-run component which is based on a Heterogeneous MIDAS polynomial structure. This uses an additive cascade of MIDAS polynomial filters moving at different frequencies in order to reproduce the changing long-run level and the persistent autocorrelation structure of high frequency trading volumes. The merits of the proposed approach are illustrated by means of an application to three stocks traded on the XETRA market characterised by different degrees of liquidity.

CO411 Room MAL 416 PERTURBATION SOLUTIONS FOR DYNAMIC ECONOMIC MODELS Chair: Peter Zadrozny

C0792: **Extended function path perturbation methods for nonstationary and unbalanced growth models**
Presenter: **Serguei Maliar**, Santa Clara University, United States
Co-authors: Vadym Lepetyuk, Lilia Maliar
Dynamic stochastic economic models are built on the assumption that the economy's fundamentals such as preferences, technologies and laws of motions for exogenous variables do not change over time. Such models have stationary solutions in which optimal value and decision functions depend on the current state but not on time. At the same time, real-world economies constantly evolve over time, experiencing population growth, technological progress, trends in tastes and habits, policy regime changes, evolution of social and political institutions, etc. There are global solution methods for solving such nonstationary and unbalanced growth models. We apply the previous methodologies to develop local perturbation methods for solving nonstationary and unbalanced growth models. Our numerical examples show that the proposed perturbation methods can deliver solutions that are comparable in accuracy to the global solutions at considerable lower cost.

C0819: **Risk-sensitive linear approximations**
Presenter: **Alexander Meyer-Gohde**, Hamburg University, Germany
Risk-sensitive approximations are constructed for policy functions of DSGE models around the stochastic steady state and ergodic mean that are linear in state variables. The method requires only the solution of linear equations using standard perturbation output to construct the approximation and is uniformly more accurate than standard linear approximations. In an application to real business cycles with recursive utility and long-run and volatility risk, the approximation successfully estimates risk aversion and stochastic volatility using the Kalman filter, where a standard linear approximation provides no information and alternative methods require computationally intensive procedures such as particle filters. At the posterior mode, the models asset pricing implications are brought in line with postwar US data without compromising the fit of the macroeconomy.

C0735: **Risk matters: Breaking certainty equivalence**
Presenter: **Juan Carlos Parra-Alvarez**, Aarhus University, Denmark
Co-authors: Olaf Posch, Hamza Polattimur
The aim is to compare the effects of uncertainty in the solution to an otherwise standard neoclassical macroeconomic model subject to technology shocks for different degrees of approximation. Our results show that certainty equivalence breaks in a continuous time version of the model even to a first order approximation, in contrast to its discrete-time version. We compare both local and global numerical methods to compute the rational expectation equilibrium dynamics and impulse response functions. We show how perturbation and collocation methods based on the Hamilton Jacobi Bellman (HJB) equation can be used to compute the models equilibrium in the space of states, fully accounting for the effects of nonlinearities and uncertainty. We also show how a first order approximation is able to capture the effects of uncertainty if risk matters. We further illustrate our results in a model known to generate substantial risk premia: the capital adjustment cost and habit formation model.

C0802: **Multi-step perturbation solution of stochastic nonlinear rational expectations models**
Presenter: **Peter Zadrozny**, Bureau of Labor Statistics, United States
Co-authors: Baoline Chen
Stochastic nonlinear rational expectations models (SNREM) are standard in macroeconomic analysis. Standard single-step perturbation (SSP) has attracted attention as a quick method for solving SNREMs with interior solutions and equations differentiable any desired number of times. SSP is a Taylor approximation of order k of an unknown solution function evaluated at point x and centered at point x_0, which is easily and accurately computed as a steady-state point. However, SSP has only local accuracy of order $|x-x_0|$ of $(k+1)$ that can be improved by increasing k at increasing costs of deriving and programming. SSP solutions are often insufficiently accurate. SSP moves from x_0 to x in one big step; MSP extends SSP by moving recursively from x_0 to x in many (h) small steps. Taylors theorem implies MSP has global accuracy of order h of $-k$, so that given x_0 and k, MSPs accuracy can be improved simply by increasing h and passing more times through an already programmed loop. 2-dimensional representations of matrix derivatives are used, which are more easily derived, comprehended, and programmed as conventional matrix equations. Equations for computing 4th-order MSP solutions for general SNREMs are derived and the method is applied to a one-sector optimal-growth model.

CO516 Room MAL 421 LABOR MARKET IMPACT OF REFUGEE IMMIGRANTS IN THE NORDIC COUNTRIES Chair: Christopher Baum

C0924: **Employment assimilation of immigrant groups in Sweden: Longitudinal evidence from Swedish register data**
Presenter: **Pieter Bevelander**, Malmo University, MIM, Sweden
Despite having one of the most celebrated labor market integration policies, the native-immigrant employment gap in Sweden is considerable, even compared to other OECD countries. We use longitudinal Swedish register data with detailed information on the reason for residence in Sweden to compare different immigrant assimilation patterns. By doing so, we are able to distinguish Refugees, Family Reunion and Labor migrants. In line with related works and theoretical considerations, we find sizable heterogeneity according to overall mean group characteristics, outmigration and employment assimilation over time. Moreover, maybe more surprisingly, our results indicate substantial differences within the group of refugees. Swedens main refugee groups differ markedly with respect to characteristics, outmigration and assimilation. The need to consider migration motivation and context in integration research and policy-making is highlighted.

C1249: **Labor market integration of refugee immigrants**
Presenter: **Bernt Bratsberg**, Frisch Center, University of Oslo, Norway

Labor market integration of refugees is examined compared to that of other immigrant groups. The study draws on Norwegian longitudinal administrative data covering labor earnings and social insurance claims over a 25-year period and presents a comprehensive picture of immigrant-native employment and social insurance differentials by admission class and by years since migration. For refugees and family immigrants from low-income source countries, we uncover encouraging signs of labor market integration during an initial period upon admission, but after just 5-10 years, the integration process goes into reverse with widening immigrant-native employment differentials and rising rates of immigrant social insurance dependency. Yet, the analysis reveals substantial heterogeneity within admission class and points to an important role of host-country schooling for successful immigrant labor market integration.

C1233: **The economic integration of immigrants in Finland: A register data analysis 2005-2014**
Presenter: **Erling Solheim**, City of Helsinki, Finland
Co-authors: Anu Yijala

The aim is to study the economic integration of immigrants in Finland, with a focus on the years 2005 to 2014. Secondary aims are to study the difference in economic integration before and after the start of the Great Recession, and to what extent it is an advantage to live and seek employment in rural communities, especially compared with the Helsinki metropolitan area. The Statistics Finland register data base FLEED with some project specific variables is the data source.

C1296: **Labour market integration of refugee-immigrants in Sweden**
Presenter: **Hans Loof**, Royal Institute of Technology, Sweden
Co-authors: Christopher Baum, Andreas Stephan, Cindy Alder

Employment probabilities and earnings differentials of refugee immigrants and natives in Sweden are assessed by following their labour market career over the period 1994-2014. Our focus is labour market integration of refugees arriving in Sweden in 1993 and 1994, mainly Balkan refugees. We study the 19541976 birth cohort, as these individuals are of working age (minimum 18 years old and maximum 60 years old) during the whole observed period. Our study contains the entire population of the target groups. Using employer-employee data, we are able to follow unique individuals accounting for time varying characteristics such as education, occupation, entrepreneurship, inventor engagement, place of living, firm size, firm productivity, firm innovation and industry. The existence of permanent long-run employment and wage gaps between refugees and natives is hypothesized. The possible gaps are tested to be explained by some systematic discrimination on the labour market, differences between refugees and natives regarding geographical mobility and occupation mobility, or to reflect differences in ability, all else equal. A large number of controls is used and a static and dynamic panel data model accounting for selectivity, endogeneity and state dependency is applied.

CO547 Room MAL 532 SEMI- AND NONPARAMETRIC METHODS FOR NONLINEAR REGRESSION Chair: Harry Haupt

C0188: **Improved local quantile regression**
Presenter: **Keming Yu**, Brunel University, United Kingdom

A new kernel-weighted likelihood smoothing quantile regression method is investigated. The likelihood is based on a normal scale-mixture representation of asymmetric Laplace distribution (ALD). This approach enjoys the same good design adaptation as the local quantile regression, particularly, for smoothing extreme quantile curves, and it ensures non-crossing quantile curves for any given sample. An application for the connections among film characteristics from the Internet Movie Database (IMDb) is promising.

C1194: **Nonlinear spatial hedonic quantile regression: Housing prices, relevant characteristics, and their shadow prices**
Presenter: **Markus Fritsch**, University of Passau, Germany
Co-authors: Harry Haupt, Joachim Schnurbus

In many applications of statistical real estate appraisal methods the following challenges arise simultaneously: [1] relevant characteristics of a property need to be identified, [2] shadow prices (marginal market valuation) of characteristics and [3] prices of bundles (of characteristics) not observed need to be estimated. State of the art hedonic housing price analysis comprises [i] modeling price functions nonlinearly, [ii] accounting for complex spatial association structures (horizontal market segmentation), and [iii] allowing for varying functional relationships across the conditional price distribution (vertical market segmentation). We discuss two general classes of nonlinear quantile regression models which meet these criteria but pursue different avenues to simultaneously address the challenges [1]-[3]. Due to the underlying assumptions, the inference obtained from both model classes differs analytically and – more importantly – leads to different economic interpretations. The methods are illustrated by applying them to data generating processes with various degrees of functional and spatial complexity in a Monte Carlo study and to geo-referenced urban house price data.

C1171: **Model confidence sets for nonparametric time series models**
Presenter: **Rolf Tschernig**, Universitaet Regensburg, Germany
Co-authors: Christoph Rust

Nonparametric estimation of nonlinear autoregressive models requires an appropriate bandwidth choice and lag selection. This can be achieved by using cross-validation. Alternatively, plug-in-bandwidth can be combined with lag selection criteria well suited for nonparametric time series models. We propose another approach: Associate for a given set of lags a specific bandwidth choice with a specific model. Then different bandwidths imply different models and bandwidth selection corresponds to model selection. This allows us to use model confidence sets. Such sets estimate the set of superior models that contains all those models that exhibit identical and lowest risk w.r.t. a user specified loss function such as quadratic loss. In practice model confidence sets typically contain several models. We suggest applying all models in the model confidence set for the modeling purpose, e.g. prediction. In a next step one can combine the model confidence sets for each lag combination to a new set and estimate a new model confidence set such that lag selection is included as well. In a Monte Carlo study we compare our proposal to existing methods.

C1773: **Modeling and forecasting nonlinear seasonality**
Presenter: **Harry Haupt**, University of Passau, Germany
Co-authors: Joachim Schnurbus

A nonparametric approach to modeling and forecasting multiple time series regressions with nonlinear seasonality is presented. A new kernel function allows us to flexibly take into account the specific periodic structure of seasonal effects. Smoothing parameters are estimated by a novel approach tailored to improve the kernel regression forecasting performance. The forecasting performance is discussed in an extensive Monte Carlo analysis and real data applications.

CO479 Room MAL 538 ADVANCES IN BEHAVIORAL MACRO-FINANCE Chair: Christian Proano

C0413: **Cross-border banking and macroprudential policies in asymmetric monetary unions**
Presenter: **Christian Proano**, University of Bamberg, Germany
Co-authors: Lena Draeger
Against the background of the emergence of macroeconomic imbalances within the European Monetary Union (EMU), we investigate the macroeconomic consequences of cross-border banking in monetary unions such as the Euro area. For this purpose, we incorporate a union-wide banking sector along previous lines in an otherwise standard two-region monetary union DSGE model, accounting for borrowing constraints of entrepreneurs and impatient households and an internal constraint on the bank's leverage ratio. We illustrate in particular how rule-of-thumb lending standards based on the macroeconomic performance of the core region within the monetary union can translate into destabilizing spill-over effects into the other region, resulting in an overall higher macroeconomic volatility. Thereby, we demonstrate a channel through which the financial sector may have exacerbated the emergence of macroeconomic imbalances within the EMU. This effect may be mitigated by macroprudential policies, where especially policies that force the bank's lending standards to be less procyclical prove to be effective in stabilizing both output in the regions of the monetary union and the leverage ratio of the financial sector.

C1174: **Modeling credit market interactions with securitization in an agent based stock flow consistent approach**
Presenter: **Thomas Theobald**, Macroeconomic Policy Institute, Germany
An agent-based stock-flow consistent model is developed with an endogenously evolving firm-bank credit network. We focus on the credit market and take into consideration the securitization process through which corporate loans of the banks' balance sheets are converted to tradable bonds by selling a pooled tranche to a special purpose vehicle (SPV). Simulation runs show that the SPV may produce losses if a significant amount of firms go bankrupt. When the SPV collapses, the banks' equity ratios deteriorate by a significant amount that might raise the entire systemic risk and lead to bank insolvencies. This effect becomes greater the higher the proportion of securitized loans which in turn might affect real economic growth through frictions on the credit market.

C0571: **Optimal constrained interest-rate rules under heterogeneous expectations**
Presenter: **Emanuel Gasteiger**, Freie Universitat Berlin, Germany
The welfare consequences of heterogeneous expectations under optimal monetary policy are examined. To this end we develop a stochastic New Keynesian model with coexistence of one-step ahead rational and adaptive expectations. We argue that the incorporation of heterogeneous expectations in both the design and implementation of optimal monetary policy is welfare improving. Nevertheless, heterogeneous expectations imply an amplification mechanism that has many adverse consequences missing under the paradigm of homogeneous rational expectations. A more hawkish policy can be welfare improving. Credible commitment eliminates or mitigates many of the ramifications of heterogeneous expectations.

C1049: **Optimal monetary policy in a mixed-frequency new Keynesian macroeconomic model with animal spirits**
Presenter: **Matthias Lengnick**, University of Kiel, FinMin SH, Germany
Co-authors: Christian Proano, Naira Kotb, Hans-Werner Wohltmann
The focus is on the difference between the Data Generating Process (DGP) of an economy and the Data Collecting Process (DCP) and the resulting implication for the design of monetary policy. Building on previous work, we assume that the financial markets activities (transactions, prices and returns) are observable at a much higher frequency (e.g. daily) than real sector activities, which are observable at a lower frequency (e.g. quarterly). This asynchrony – present in the real world but widely ignored in macroeconomic theory – posits interesting questions both in terms of modeling and in terms of optimal monetary policy design. We develop a behavioral New Keynesian Model with an integrated financial market where all dynamics (real sector & financial sector) are defined on daily intervals. Real sector dynamics are observable by the agents only quarterly. However, the central bank is assumed superior to the agents in the sense that it has the capability of observing the dynamics, both in the real and the financial sector, on daily basis. Optimal monetary policy analysis is conducted, through which different linear and nonlinear policy rules are compared. The main questions are, is it worth it for the monetary policy to react to daily dynamics? And when should monetary policy react to financial data?

CO320 Room MAL 539 TIME-SERIES ECONOMETRICS Chair: Robert Kunst

C0446: **Using a global vector autoregression to conditionally forecast tourism exports and tourism export prices**
Presenter: **Ulrich Gunter**, MODUL University Vienna, Austria
The purpose is to analyze the ex-ante projected trajectories of real tourism exports and relative tourism export prices of the EU-15 conditional on expert real GDP growth forecasts for the global economy that were provided by the OECD for the years 2013 to 2017. To this end, a global vector autoregression (GVAR) is applied to a panel data set ranging from 1994Q1 to 2013Q3 for a cross section of 45 countries. In line with economic theory, growing global tourist income combined with decreasing relative destination price ensures, in general, increasing tourism demand for the politically and macroeconomically distressed EU-15. However, the conditionally forecast increases in tourism demand are under-proportional for some EU-15 member countries. Thus, rather than simply relying on increases in tourist income at the global level, tourism planners and developers in the EU-15 should address the relatively low price-competitiveness of the member countries in order to counter the rising competition for global market shares and to ensure future tourism export earnings.

C0743: **Simulation-based selection of prediction models**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
The benefits of basing model selection decisions in a forecasting context are assessed on simulations that fuse data information and the structure hypothesized by tentative rival models. These procedures can be applied to any empirical forecasting problems. The focus is, however, on macroeconomic applications. The suggested procedure aims at choosing among a small number of tentative forecast models in the presence of data. From models fitted to the data, pseudo-data are generated. Again, the models are applied to the pseudo-data and their out-of-sample performance is evaluated. The ultimate choice of the forecasting model is based on the relative performance of rival models in predicting "their own data" and those of the rival model. The project covers the three aspects of a rigorous statistical foundation, of a Monte Carlo evaluation of the procedure, and of exemplary empirical implementations of the method in typical macro-economic applications.

C1806: **Scale-invariant CUSUM-based ratio tests for parameter constancy: Application to variance stability**
Presenter: **Paulo Rodrigues**, Universidade Nova de Lisboa, Portugal
New CUSUM based tests for parameter constancy are introduced following a class of ratio tests recently proposed, which are based on a Karhunen-Loeve expansion. Theoretical results are presented and an in-depth Monte Carlo analysis is performed to evaluate the finite sample performance of the procedure. A comparison with available approaches is provided. The new tests present superior finite sample size properties and interesting power behaviour. An empirical application to the Bitcoin Coindesk index is also included.

C0488: **Stable limits for the Gaussian QMLE in the non-stationary GARCH(1,1) model**
Presenter: **Stelios Arvanitis**, RC-AUEB, Greece
Co-authors: Alexandros Louka
The limit theory of the Gaussian QMLE in the non-stationary GARCH(1,1) model is derived when the squared innovation process lies in the domain of attraction of a stable law. Analogously to the stationary case, when the stability parameter lies in $(1,2]$, we find regularly varying rates and stable limits for the QMLE of the ARCH and GARCH parameters.

CO226 Room MAL 540 COLLINEARITY, COMMON FACTORS AND COMMON SHOCKS Chair: Niklas Ahlgren

C0312: **Stargazing with structural VARs: Shock identification via independent component analysis**
Presenter: **Aleksei Netsunajev**, Tallinn University of Technology, Estonia
Co-authors: Dmitry Kulikov
A new ICA-based statistical identification procedure is introduced for Bayesian SVAR models with independent non-Gaussian structural innovations. Additional statistical information, available in non-normally distributed shocks, allows us to estimate a fully general structural VAR model without use of any strong conventional a priori identifying restrictions. The new procedure is validated using the US macroeconomic data series, where the nature of four different shocks is empirically examined. In particular, we statistically identify the short-run impacts and impulse responses of four structural shocks, which we label a monetary policy shock, a money demand shock, an aggregate demand shock and an aggregate supply shock. We find a robust and well-pronounced price puzzle in response to our monetary policy shock, while the statistically identified money demand shock induces a strong reaction in the US real output and prices. In addition, we also document a new real output puzzle in response to the Federal Reserve monetary policy action.

C0547: **Strong and weak cross-sectional dependence in factor models for returns in event studies**
Presenter: **Niklas Ahlgren**, Hanken School of Economics, Finland
Co-authors: Jan Antell
Cross-sectional dependence in abnormal returns is investigated using a recently introduced exponent of cross-sectional dependence. The abnormal return is defined as the actual return minus the normal return predicted by a factor model. The factor model is valid for computing abnormal returns provided there is no remaining factor structure (strong cross-sectional dependence), so that abnormal returns are at most weakly cross-sectionally dependent. We compare five commonly used factor models for returns. The models are the market model, three-factor model, three-factor model extended with momentum factor, five-factor model and five-factor model extended with momentum factor. The models are fitted to US stock returns from Bear Stearns' collapse and Lehman Brothers' bankruptcy in 2008. These data provide evidence that the cross-sectional dependence in returns due to common factors other than the market factor is strong. Contrary to recommendations in the literature to use the market model, we propose the use of multi-factor models to compute abnormal returns in event studies.

C0537: **Testing collinearity of vector time series**
Presenter: **Agnieszka Jach**, Hanken School of Economics, Finland
Co-authors: Tucker McElroy
The collinearity of vector time series in the frequency domain is investigated by examining the rank of the spectral density matrix at a given frequency of interest. Rank reduction corresponds to collinearity at the given frequency. When the time series data is nonstationary and has been differenced to stationarity, collinearity corresponds to co-integration at a particular frequency. We pursue a full understanding of rank through the Schur complements of the spectral density matrix, and test for rank reduction via assessing the positivity of these Schur complements, which are obtained from a nonparametric estimator of the spectral density. We provide new asymptotic results for the Schur complements, under the fixed bandwidth ratio paradigm. The test statistics are $O_P(1)$ under the alternative, but under the null hypothesis of collinearity the test statistics are $O_P(T^{-1})$, and the limiting distribution is non-standard. Subsampling is used to obtain the limiting null quantiles. Simulation study and an empirical illustration for six-variate time series data are provided.

C0593: **Taking zero lower bound seriously: A structural vector autoregression containing positive-valued components**
Presenter: **Henri Nyberg**, University of Turku and University of Helsinki, Finland
In the conventional structural vector autoregressive (SVAR) analysis, the variables are treated as real-valued components. This is not, however, always the case as seen, for example, during the recent Zero Lower Bound state in nominal short-term interest rates. A SVAR model containing at least one strictly positive component is developed. Despite the nonlinearity originating from the positive-valued components, impulse response functions and forecast error variance decompositions can be expressed with explicit analytical formulae. The empirical results of the core three-variable U.S. monetary policy system generally point out important differences to the linear SVAR models.

CO453 Room MAL 541 ROBUST BAYESIAN ECONOMETRICS Chair: Catherine Forbes

C1441: **Monitoring economic linkage with a semi-parametric VAR model**
Presenter: **Hong Wang**, Monash University, Australia
Co-authors: Catherine Forbes, Bonsoo Koo
An extension to the Global Vector AutoRegressive (Global-VAR) model for economic variables observed with mixed frequency is proposed. The modelling framework employs the notion that certain variables observed only at low frequency (e.g. quarterly) have latent (unobserved) high frequency (e.g. monthly) values that are "missing". Such variables are nevertheless related to several high frequency observed variables through a regression framework. Bayesian analysis of the model, which is specified in hierarchical form, is amenable to the use of Markov chain Monte Carlo methods. Two cases are considered, where the latent variable evolves linearly, and where the latent variable evolves nonlinearly, with the latter specified non-parametrically using a Gaussian Process prior. The proposed modeling framework is applied to progressively to "nowcast" (i.e. predict the present state of) the missing values and their corresponding quarterly aggregates. Data available from 36 countries, and spanning the period from 1979:Q2 to 2013:Q4, are explored, with the nowcast distributions resulting from the two approaches evaluated over an (expanding) out-of-sample period that includes the Global Financial Crisis.

C1415: **A Bayesian quantile time series model for asset returns**
Presenter: **Gelly Mitrodima**, LSE, United Kingdom
Co-authors: Jim Griffin
The conditional distribution of asset returns has been widely studied in the literature using a wide range of methods that usually model its conditional variance. However, empirical studies show that other features of the distribution may also vary over time. In particular, the returns of most assets display time-dependence beyond volatility, and there is difficulty with fitting their extreme tails. Our aim is to study the time variation in the shape of the return distribution described by a collection of conditional quantiles. Direct modelling of quantile for Bayesian inference is challenging, since it involves analytic expressions for both the quantile function and its inverse to define the likelihood. Thus, we propose a novel class of Bayesian nonparametric priors for quantiles built around a random transformation. This allows fast and efficient Markov chain Monte

Carlo (MCMC) methods to be applied for posterior simulation and forecasting. Under this Bayesian nonparametric framework, we avoid strong parametric assumptions about the underlying distribution, and so we obtain a model that is flexible about the shape of the distribution. We define a stationary model and we derive the stationary mean and variance of the quantiles. In our empirical exercise, we find that the model fits the data well, offers robust results, and acceptable forecasts for a sample of stock, index, and commodity returns.

C0845: **Robust Bayesian inference for moment condition models**
Presenter: **Catherine Forbes**, Monash University, Australia
Co-authors: Zhichao Liu, Heather Anderson

A new robust Bayesian exponentially tilted empirical likelihood (RBETEL) inferential methodology is proposed which is suitable for moment condition models when data may be contaminated by outliers. It is built upon the Bayesian exponentially tilted empirical likelihood (BETEL) method, justified by the fact that an empirical likelihood (EL) can be interpreted as the nonparametric limit of a Bayesian procedure when the implied probabilities are obtained from maximizing entropy subject to some given moment constraints. The BETEL method is found to be linked to a general framework which updates prior belief via a loss function. After demonstrating that the BETEL loss function is related to the EL ratio, a loss function for the new RBETEL method arises naturally as the EL ratio evaluated on some sub-samples. The resulting posterior distribution for the parameters is shown to be a coherent representation of the subjective uncertainty in the minimizer of the expected loss. A controlled simulation experiment is conducted to investigate the performance of the RBETEL method. We find that the proposed methodology produces reliable posterior inference for the fundamental relationships that are embedded in the majority of the data, even when outliers are present in the dataset. The method is also illustrated in an empirical study relating inflation and economic openness.

CG325 Room MAL 152 CONTRIBUTIONS IN STRUCTURAL VAR Chair: Luca Fanelli

C1454: **Identification of SVAR models by combining sign restrictions with external instruments**
Presenter: **Robin Braun**, University of Konstanz, Germany
Co-authors: Ralf Brueggemann

Structural vector autoregressive (SVAR) models are identified by combining sign restrictions with information in external instruments and proxy variables. We incorporate the proxy variables by augmenting the SVAR with equations that relate them to the structural shocks. Our modeling framework allows to simultaneously identify different shocks using either sign restrictions or an external instrument approach, always ensuring that all shocks are orthogonal. The combination of restrictions can also be used to identify a single shock. This entails discarding models that imply structural shocks that have no close relation to the external proxy time series, which narrows down the set of admissible models. Our approach nests the pure sign restriction case and the pure external instrument variable case. We discuss full Bayesian inference, which accounts for both, model and estimation uncertainty. We illustrate the usefulness of our method in SVARs analyzing oil market and monetary policy shocks. Our results suggest that combining sign restrictions with proxy variable information is a promising way to sharpen results from SVAR models.

C1563: **Uncertainty across volatility regimes**
Presenter: **Luca Fanelli**, University of Bologna, Italy
Co-authors: Giovanni Angelini, Emanuele Bacchiocchi

The aim is to employ a small-scale non-recursive structural vector autoregression (SVAR) and a new identification methodology to address empirically two questions by using post-WW2U.S. data: (i) does the relationship between macroeconomic and financial uncertainty and economic activity change across main macroeconomic regimes? (ii) Is macroeconomic and/or financial uncertainty a major cause or effect (or both) of decline in economic activity? Our non-recursive SVAR is identified by exploiting the breaks in the (unconditional) volatility that characterize macroeconomic variables across business cycle and generates regime-dependent impulse response functions. To address (ii), the specified system features a number of moment conditions induced by heteroscedasticity which permit modeling 'reverse causality', i.e. uncertainty shocks can cause a decline in real economic activity and, simultaneously, heightened uncertainty can represent a contemporaneous response to real economic activity shocks. Empirical results based on monthly data on the period 1960-2015 suggest a positive answer to (i) and a more involved answer to (ii). Our evidence also suggests that during the zero lower bound on the federal fund rate, the implementation of unconventional policies by the Fed possibly contributed to attenuate the mechanisms behind the rise of endogenous volatility/uncertainty.

C1511: **GMM estimation of structural vector autoregressions**
Presenter: **Markku Lanne**, University of Helsinki, Finland
Co-authors: Jani Luoto

A structural vector autoregressive (SVAR) model with an independent non-Gaussian error vector is known to be locally identified without any further restrictions. However, its estimation by the method of maximum likelihood entertained in the previous literature requires the specification of a particular non-Gaussian error distribution. Moreover, for statistical inference, additional restrictions are needed to ensure global identification. We propose a generalized method of moments (GMM) estimator for the SVAR model that avoids explicit distributional assumptions by making use of independence and non-Gaussian features of the structural shocks. Independence implies no contemporaneous correlation and no lead-lag relations between the components of the error vector, while non-Gaussianity is captured by moment conditions related to their co-kurtosis. When the moment conditions only involve lead-lag relations in one direction, or co-kurtosis is defined using skewness terms, identification up to the signs of the structural shocks is achieved. Commonly used moment selection criteria are applicable in our setup, and tests of over-identifying restrictions can be used to assess the adequacy of identification. According to simulation results, two-step GMM estimation works well even in small samples, and is robust with respect to the moments selected. An empirical application to postwar U.S. macroeconomic data illustrates the methods.

C1643: **Identification of independent structural shocks in the presence of multiple Gaussian components**
Presenter: **Simone Maxand**, University of Goettingen, Germany

Several recently developed identification techniques for structural VAR models are based on the assumption of non-Gaussianity. So-called independence based identification provides unique structural shocks (up to scaling and ordering) under the assumption of at most one Gaussian component. While non-Gaussianity of certain interesting shocks, e.g., a monetary policy shock, appears rather natural, not all macroeconomic shocks in the system might show this clear difference from Gaussianity. We generalize identifiability by noting that even in the presence of multiple Gaussian shocks the non-Gaussian ones are still unique. Consequently, independence based identification allows to uniquely determine the (non-Gaussian) shocks of interest irrespective of the distribution of the remaining system. In an illustrative macroeconomic model the identified structural shocks confirm the results of previous studies on the early millennium slowdown. Furthermore, extending the time horizon provides full identification under the non-Gaussianity assumption.

CG119 Room MAL 402 CONTRIBUTIONS IN LONG MEMORY Chair: Michelle Voges

C0218: **Optimal bias-correction in the log periodogram estimation of the fractional parameter: A jackknife approach**

Presenter: **Kanchana Nadarajah**, Monash University, Australia, Australia

Co-authors: Gael Martin, Donald Poskitt

A bias-corrected log-periodogram regression (LPR) estimator based on a jackknife method is proposed for the fractional parameter in a stationary class of fractionally integrated models. The weights of the full sample and the subsamples in the jackknife estimator are chosen such that bias reduction occurs to an order of n^{-a}, where $0 < a < 1$ with n the sample size, without inflating the asymptotic variance. Some statistical properties of the discrete Fourier transform and periodograms related to the full sample and subsamples of a time series, which are exploited in constructing the optimal weights, are also provided. Under some regularity conditions, it is shown that the optimal jackknife estimator is consistent and has a limiting normal distribution with the same asymptotic variance and rate of convergence as the original LPR estimator. These theoretical results are valid under both non-overlapping and moving-block sub-sampling schemes used in the jackknife technique. In a Monte Carlo study the new optimal jackknife estimator outperforms some of the main competing estimators, in terms of bias-reduction and root mean-squared-error. The proposed method is applied to two empirical time series: (i) the annual minimum level of the Nile River, and (ii) realized volatility for the daily returns on the S&P500 index.

C1488: **Fractal methods for fractional cointegration**

Presenter: **Ladislav Kristoufek**, Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

Detrended fluctuation analysis (DFA) and detrending moving average (DMA) methods are standardly used for fractional differencing parameter d estimation. Recently, the DFA and DMA based estimators of standard regression parameters have been proposed. The estimators possess some desirable properties with regards to long-range dependence, trends, seasonalities and heavy tails. We study properties of both estimators beyond the general fractional cointegration framework, i.e. we examine a simple model $y_t = \alpha + \beta x_t + u_t$, where $x_t \sim I(d)$ and $u_t \sim I(d-b)$, which implies $y_t \sim I(\max[d, d-b])$. The fractional cointegration requires $b > 0$, while the standard cointegration $CI(1,1)$ assumes $x_t, y_t \sim I(1)$ and $u_t \sim I(0)$. We are interested in various combinations of d and b parameters ($0 \leq b \leq 1$, i.e. we cover not only the fractional cointegration framework). We provide a broad Monte Carlo simulation study focusing on different time series lengths, combination of d and b parameters, and on possible spurious relationships. Specifically, we compare the estimators based on DFA and DMA with the standard OLS procedure under true and spurious relationships ($\beta = 0$ and $\beta \neq 0$). Based on the bias, standard error and mean squared error of the estimators, the new procedures outperform OLS for various settings (e.g. with $d = 1$ and $b < 0.5$).

C1537: **Long horizon forecasts**

Presenter: **Jose Eduardo Vera Valdes**, Aalborg University, Denmark

Most forecasting studies for long memory assume that the series are generated by fractional processes. We assess the performance of the ARFIMA model when forecasting long memory series where the long memory generating mechanism may be different from fractional differencing. We consider cross sectional aggregation and the error duration model as long memory generating mechanisms. We find that ARFIMA models produce similar forecast performance compared to high-order AR models at shorter horizons. As the forecast horizon increases, the ARFIMA models tend to dominate in forecast performance. Hence, ARFIMA models are well suited for long horizon forecasts of long memory processes regardless of how the long memory is generated. Additionally, we analyze the forecasting performance of the heterogeneous autoregressive model (HAR), we find that the structure enforced by the HAR model produces better long horizon forecasts than AR models of the same order at the price of inferior short horizon forecasts in some cases. Our results have implications for Climate Econometrics and Financial Econometrics models dealing with larger forecast horizons. In an example, we show that a short memory ARMA model gives the best performance when forecasting the Realized Variance for the S&P 500 up to a month ahead, while an ARFIMA model gives the best performance for longer forecast horizons.

C0263: **The disintegration of EMU government bond markets**

Presenter: **Michelle Voges**, Leibniz University Hannover, Germany

Co-authors: Christian Leschinski, Philipp Sibbertsen

The aim is to analyze market integration among long term government bonds in the Eurozone from the introduction of the Euro in 1999 to the Subprime mortgage crisis and the EMU debt crisis. We test for the existence and stability of equilibria between the interest rates using semiparametric tests for fractional cointegration and also propose a test for the dissolution of an existing fractional cointegrating relationship. This econometric framework allows us to establish several new findings. We find that EMU government bond markets were economically integrated in the initial phase after the introduction of the Euro, but markets disintegrated shortly thereafter, which was long before the crisis periods. We find that for most of the time since the introduction of the Euro there was no correction mechanism that would push interest rates on EMU government bonds towards a joint equilibrium. These findings can be explained by arguments from the theory on optimum currency areas and we discuss their implications for the interrelation of the Subprime mortgage crisis and the EMU debt crisis.

Saturday 16.12.2017 16:45 - 18:50 Parallel Session F – CFE-CMStatistics

EI013 Room CLO B01 BOUNDARY ESTIMATION AND DECONVOLUTION PROBLEMS **Chair: Leopold Simar**

E0154: **A novel method for panel models with jump discontinuities**
Presenter: **Alois Kneip**, University of Bonn, Germany
While a substantial literature on structural break change point analysis exists for univariate time series, research on large panel data models has not been as extensive. A novel method for estimating panel models with multiple structural changes is proposed. The breaks are allowed to occur at unknown points in time and may affect the multivariate slope parameters individually. Our method is related to the Haar wavelet technique; we adjust it according to the structure of the observed variables in order to detect the change points of the parameters consistently. We also develop methods to address endogeneous regressors within our modeling framework. The asymptotic property of our estimator is established. In our application, we examine the impact of algorithmic trading on standard measures of market quality such as liquidity and volatility over a time period that covers the financial meltdown that began in 2007. We are able to detect jumps in regression slope parameters automatically without using ad-hoc subsample selection criteria.

E0155: **Nonparametric covariate-adjusted regression**
Presenter: **Aurore Delaigle**, University of Melbourne, Australia
Co-authors: Peter Hall, Wenxin Zhou
Nonparametric estimation of a regression curve is considered when the data are observed with multiplicative distortion which depends on an observed confounding variable. We suggest several estimators, ranging from a relatively simple one that relies on restrictive assumptions usually made in the literature, to a sophisticated piecewise approach that involves reconstructing a smooth curve from an estimator of a constant multiple of its absolute value, and which can be applied in much more general scenarios. We show that, although our nonparametric estimators are constructed from predictors of the unobserved undistorted data, they have the same first order asymptotic properties as the standard estimators that could be computed if the undistorted data were available. We illustrate the good numerical performance of our methods on both simulated and real datasets.

E0156: **Estimation of the boundary of a variable observed with symmetric error**
Presenter: **Ingrid Van Keilegom**, KU Leuven, Belgium
Co-authors: Jean-Pierre Florens, Leopold Simar
Consider the model $Y = X + \varepsilon$ with $X = \tau + Z$, where τ is an unknown constant (the boundary of X), Z is a random variable defined on R^+, ε is a symmetric error, and ε and Z are independent. Based on an iid sample of Y, we aim at identifying and estimating the boundary τ when the law of ε is unknown (apart from symmetry) and in particular its variance is unknown, contrary to most papers in the literature. We propose an estimation procedure based on a minimal distance approach and by making use of Laguerre polynomials. Asymptotic results as well as finite sample simulations are shown.

EO186 Room MAL B20 APPLIED EXTREMES **Chair: Gilles Stupfler**

E0198: **Estimating space-time trend and dependence of heavy rainfall**
Presenter: **Ana Ferreira**, Instituto Superior Tecnico, Portugal
Co-authors: Petra Friederichs, Laurens de Haan, Claudia Neves, Martin Schlather
A new approach for evaluating time-trends in extreme values accounting also for spatial dependence is proposed. Based on exceedances over a space-time threshold, estimators for a trend function and for extreme value parameters are given, leading to a homogenization procedure for then applying stationary extreme value processes. Extremal dependence over space is further evaluated through variogram analysis including anisotropy. We detect significant inhomogeneities and trends in the extremal behaviour of daily precipitation data over a time period of 84 years and from 68 observational weather stations in North-West Germany. We observe that the trend is not monotonous over time in general. Asymptotic normality of the estimators under maximum domain of attraction conditions are proven.

E1092: **Towards an efficient early warning system for extreme wind speed detection**
Presenter: **Daniela Castro**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Amanda Hering
Renewable sources of energy such as wind power have become a sustainable alternative to fossil fuels-based energy. However, the uncertainty and fluctuation of wind speeds brings a great threat to the wind power production stability, and to the wind turbines themselves. A turbine cut-off point denotes how fast the turbine can go before turbine blades are brought to rest to prevent any damage produced by extreme wind speeds. Therefore, one of the main problems related to extremes in wind is to forecast when the wind speed will exceed this cut-off speed. We develop a flexible early warning system to detect the risk of extreme winds at a given station based on a set of neighboring stations and the dominant wind regime. The main challenges with this approach are the temporal non-stationarity of the wind regimes, and the fact that wind speeds are inherently anisotropic, which implies that the set of neighboring stations can potentially be different with different wind regimes. To cope with these issues we fit a flexible spatial copula model to threshold exceedances, in order to determine the probability of observing an extreme event at a specific site, given the dynamic wind regimes and the information gathered from a set of neighboring stations. Our approach is illustrated with data measured at meteorological towers located in the Pacific Northwest, US.

E0757: **An extreme value analysis of top performing UK winter wheat producers**
Presenter: **Emily Mitchell**, The University of Nottingham, United Kingdom
Co-authors: Gilles Stupfler, Andrew Wood, Neil Crout, Paul Wilson
Using the responses to a UK-based survey, we present the first application of extreme value theory in an agricultural setting to complement the previous studies conducted from a classical central perspective in this field. The Farm Business Survey collects a substantial amount of information annually from farms across England and Wales with the purpose of providing farmers with an overview of farming performances. Winter wheat is the most popular crop grown in the UK due to its optimal growing conditions; therefore, we focus on winter wheat production from 2006 to 2015 and extract a subset of variables from this data set, among which the obtained yield and net margin, and apply a number of established extreme value analysis methods. In particular, we use a mix of Peaks-Over-Threshold and semi-parametric approaches to fit distributions to the tail before ultimately producing extreme quantile estimates. We conclude by discussing the implications of our results regarding top UK winter wheat producers, and especially how their financial results compare to these of top earners in the UK.

E1167: **Feature clustering and tests for asymptotic independence**
Presenter: **Anne Sabourin**, Telecom ParisTech, France
Co-authors: Mael Chiapino, Johan Segers
Sparsity in multivariate extremes may be defined as the concentration of the angular measure on low dimensional subsets of the unit sphere. In a

high dimensional context ($d > 50$), a natural first step when analyzing the dependence structure of extreme event is to identify such a pattern. Earlier works have proposed different algorithms for this purpose. The stopping criterion for the CLEF algorithm is shown to be a linear combination of extremal coefficients. This allows us to interpret the CLEF algorithm, up to a minor re-tuning of its stopping criterion, as a sequence of tests for multivariate asymptotic full dependence between various subsets of components, with asymptotic guarantees. We also propose deeper modifications of the CLEF algorithm obtained by replacing the null hypothesis of asymptotic dependence with that of asymptotic independence and using test statistics based on coefficients of multivariate tail dependence.

E0609: **Statistical downscaling for future extreme wave heights in the North Sea**
Presenter: **Ross Towe**, Lancaster University, United Kingdom
Co-authors: Emma Eastoe, Jonathan Tawn, Philip Jonathan

For safe offshore operations accurate knowledge of the extreme oceanographic conditions is required. We develop a multi-step statistical downscaling algorithm using data from low resolution global climate model (GCM) and local-scale hindcast data to make predictions of the extreme wave climate in the next 50 year period at locations in the North Sea. The GCM is unable to produce wave data accurately so instead we use its 3-hourly wind speed and direction data. By exploiting the relationships between wind characteristics and wave heights, a downscaling approach is developed to relate the large and local-scale data sets and hence future changes in wind characteristics can be translated into changes in extreme wave distributions. We assess the performance of the methods using within sample testing and apply the method to derive future design levels over the northern North Sea.

EO051 Room MAL B33 HIGH DIMENSIONAL MODEL SELECTION Chair: Malgorzata Bogdan

E0340: **A novel algorithmic approach to Bayesian logic regression**
Presenter: **Florian Frommlet**, Medical University Vienna, Austria

Logic regression was developed more than a decade ago as a tool to construct predictors from Boolean combinations of binary covariates. It has been mainly used to model epistatic effects in genetic association studies, which is very appealing due to the intuitive interpretation of logic expressions to describe the interaction between genetic variations. Nevertheless logic regression has remained less well known than other approaches to epistatic association mapping. The aim is to introduce an advanced evolutionary algorithm called GMJMCMC (Genetically modified Mode Jumping Markov Chain Monte Carlo) to perform Bayesian model selection in the space of logic regression models. Comprehensive simulation studies illustrate its performance given logic regression terms of various complexity. Specifically GMJMCMC is shown to be able to identify three-way and even four-way interactions with relatively large power, a level of complexity which has not been achieved by previous implementations of logic regression. GMJMCMC is applied to reanalyze QTL mapping data for Recombinant Inbred Lines in Arabidopsis thaliana and for a backcross population in Drosophila.

E1165: **Using tensor regression to select the HIV-related connections between brain's regions**
Presenter: **Damian Brzyski**, Indiana University Bloomington, United States
Co-authors: Joaquin Goni, Jaroslaw Harezlak, Beau Ances

Classical regression methods treat covariates as a vector and estimate a corresponding vector of regression coefficients. In medical applications, however, very often regressors in a form of multidimensional arrays could be met. For instance, one might be interested in utilizing 2- or 3-dimensional MRI brain images to determine brain regions whose structural or functional features are associated with certain medical condition. Flattening such image array into a vector is an unsatisfactory solution, since it destroys the inherent spatial structure of the image and could be very challenging from the computational point of view. We will present the alternative approach - the tensor regression - which we are using to investigate the changes in the structure of brain's connections in the cohort of HIV-positive subjects. For each patient such connectivity information occurs in the form of p by p dimensional matrix, where p is a number of considered brain's regions and where entries indicate how strongly pairs of regions are connected. The corresponding model's coefficients are in the form of p by p dimensional matrix as well and the goal is to select the response-related entries.

E1188: **Sorted-L1 norm for outliers detection and high-dimensional robust regression: Sharp oracle inequalities and FDR control**
Presenter: **Alain Virouleau**, Ecole polytechnique, France
Co-authors: Stephane Gaiffas, Agathe Guilloux, Malgorzata Bogdan

The problems of outlier detection and robust regression in a high-dimensional setting are fundamental in statistics, and have numerous applications. Following a recent set of works providing methods for simultaneous robust regression and outliers detection, we consider a model of linear regression with individual intercepts, in a high-dimensional setting. Each individual intercept, if non-zero, corresponds to an individual shift from the linear model. In this setting, we introduce a new procedure for simultaneous estimation of the linear regression coefficients and intercepts, using two dedicated sorted-L1 penalizations, also called SLOPE. We develop a complete theory for this problem: first, we provide sharp oracle inequalities on the statistical estimation error of both the vector of individual intercepts and regression coefficients.Second, we give an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the intercepts, namely for the problem of outliers detection, obtained through our method. It is noteworthy that this is the first attempt to introduce a procedure for this problem with guaranteed FDR and statistical power control. Numerical illustrations, with a comparison to recent alternative approaches, are provided on both simulated datasets and several real-world datasets.

E0886: **Improving Lasso for sparse high dimensional GLM and Cox model selection**
Presenter: **Piotr Pokarowski**, University of Warsaw, Poland
Co-authors: Agnieszka Prochenka, Michal Frej, Jan Mielniczuk

The Lasso, that is l_1-penalized loss estimator is a very popular tool for fitting sparse high dimensional models. However, theory and simulations established that the model selected by the Lasso is usually too large. The concave regularizations (SCAD, MCP or capped-l_1) are closer to l_0-penalized loss, that is, to the Generalized Information Criterion (GIC) than the Lasso, and they correct its intrinsic estimation bias. That methods use the Lasso as a starting set of models and try to improve it using local optimization. We propose an alternative method of improving the Lasso for Generalized Linear Models and Cox Models which is a generalization of our SOS algorithm for linear models. The method, for a given penalty, orders the absolute values of the Lasso non-zero coordinates and then selects the model from a small family by GIC. We derive upper bounds on the selection error of the algorithm and show in numerical experiments on synthetic and real-world data sets that an implementation of our algorithm is more accurate than implementations of concave regularizations.

E1038: **Gene hunting with knockoffs for hidden Markov Models**
Presenter: **Matteo Sesia**, Stanford University, United States
Co-authors: Chiara Sabatti, Emmanuel Candes

Modern scientific studies often require the identification of a subset of relevant explanatory variables, in the attempt to understand an interesting phenomenon. Several statistical methods have been developed to automate this task, but only recently has the framework of model-free knockoffs proposed a general solution that can perform variable selection under rigorous type-I error control, without relying on strong modeling assumptions.

We extend the methodology of model-free knockoffs to a rich family of problems where the distribution of the covariates can be described by a hidden Markov model (HMM). We develop an exact and efficient algorithm to sample knockoff copies of an HMM. We then argue that combined with the knockoffs selective framework, they provide a natural and powerful tool for performing principled inference in genome-wide association studies with guaranteed false discovery rate control. Finally, we apply our methodology to several datasets aimed at studying Crohn's disease and several continuous phenotypes, e.g. levels of cholesterol.

EO065 Room MAL B34 STATISTICS FOR HILBERT SPACES Chair: Gil Gonzalez-Rodriguez

E0321: **Nonparametric variable selection and screening with a large number of functional predictors**
Presenter: **Jan Gertheiss**, Clausthal University of Technology, Germany
The situation where a very large number, like thousands, or tens of thousands, of functional predictors are available in a regression or classification problem is considered. A typical example are time course gene expression data used for classifying patients into different groups. Standard approaches for variable (pre)selection in this context are typically based on univariate, ANOVA-type testing, followed by ranking covariates according to the associated p-values. We propose an alternative, nonparametric method: the functional nearest neighbor ensemble. Nearest neighbor based estimates are used to build an ensemble, with each ensemble member representing a specific functional covariate. A lasso-type penalty is used to select ensemble members, and hence functional predictors with the highest predictive power. The proposed ensemble method is illustrated in simulation studies and on real world data, and compared to standard approaches described above.

E0903: **Doubly functional graphical models in high dimensions**
Presenter: **Gareth James**, University of Southern California, United States
Co-authors: Xinghao Qiao, Cheng Qian
The problem of estimating a functional graphical model from a data set consisting of functional observations is considered. Recent work in this area has focused on modelling dynamic graphs from high dimensional but time-varying-distributed scalar data. However, many real world examples require the construction of networks for multivariate functional data. We present a novel perspective by treating dynamic edge sets, which, for Gaussian data, correspond to dynamic sparse precision matrices, as functional objects, linking the concept of dynamics with multivariate functional data. A class of doubly functional graphical models is proposed to capture this evolving conditional dependence relationship among multiple random functions. Our approach first estimates a functional covariance matrix, and then computes sparse precision matrices, which in turn provide the doubly functional graphical model.

E1056: **High-dimensional function-on-scalar regression in Hilbert spaces**
Presenter: **Matthew Reimherr**, Pennsylvania State University, United States
Co-authors: Alice Parodi
Recent work concerning function-on-scalar regression when the number of predictors is much larger than the sample size is discussed. In particular, we will present a new methodology, called FLAME for Functional Linear Adaptive Mixed Estimation, which simultaneously selects, estimates, and smooths the important predictors in the model. Our methodology is readily available as an R package that utilizes a coordinate descent algorithm for fast implementation. Asymptotic theory will be provided and we will compare to previous methods via simulations.

E0973: **Quadratic regression for functional response models**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan
The problem of constructing a regression model with a functional predictor and a functional response is considered. We extend the functional linear model to the quadratic model, where the quadratic term also takes the interaction between the argument of the functional data into consideration. We assume that the predictor and the coefficient functions are expressed by basis expansions, and then parameters included in the model are estimated by the maximum likelihood method by assuming that the error function follows a Gaussian process. We apply the proposed method to the analysis of weather data, and then investigate what the results provides.

E0890: **On the optimal reconstruction of partially observed functional data**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Alois Kneip
A new reconstruction operator is proposed which aims to recover the missing parts of a function given the observed parts. The new reconstruction operator belongs to a new large class of functional operators which includes the classical regression operators, considered so far, as a special case. We show the optimality of our reconstruction operator among this new class of operators. Our estimation theory allows for autocorrelated functional data and considers the practically relevant situation where each of the n functions is observed at m discretization points. We derive uniform rates of consistency for our nonparametric estimation procedures using a double asymptotic that allows investigate all data scenarios from almost sparse to dense functional data. The finite sample properties are investigated through simulations and a real data application.

EO425 Room MAL B35 RECENT ADVANCES IN STATISTICAL MACHINE LEARNING Chair: Peter Radchenko

E0952: **High-dimensional regression with L0 regularization**
Presenter: **Peter Radchenko**, University of Sydney, Australia
Co-authors: Antoine Dedieu, Rahul Mazumder
New applications of discrete optimization techniques in high-dimensional regression will be discussed. In particular, we will review the recently proposed mixed integer optimization implementations of the best subset selection estimator and the discrete Dantzig Selector. The latter estimator minimizes the number of nonzero regression coefficients, subject to a budget on the maximal absolute correlation between the features and residuals. It can be expressed as a solution to a mixed integer linear optimization problem, a computationally tractable framework that delivers provably optimal global solutions. The current state of algorithmics in integer optimization makes the estimator highly scalable: it scales gracefully to problems with 10,000 predictors. We will also discuss applications of mixed integer optimization in high-dimensional linear regression with group structure, as well as high-dimensional additive modeling. In addition, we will consider a regularized version of the best subset selector, and investigate its advantages in the low signal regimes.

E1258: **Fast L0 regression: Coordinate descent, local search, and combinatorial optimization**
Presenter: **Hussein Hazimeh**, Massachusetts Institute of Technology, United States
Co-authors: Rahul Mazumder
The computation of estimators for the L_0 penalized regression problem with possible additive regularizers is considered. While these estimators exhibit nice statistical properties, current algorithms for this class of problems can be slow when compared to the efficient algorithms for the Lasso (e.g., glmnet). The aim is to reduce this gap. We explore coordinate descent-type methods and show (both in theory and practice) that they outperform the commonly used iterative hard-thresholding type methods in terms of solution quality and computation times. We develop a unified framework in which local search methods, pathwise optimization, and mixed integer optimization are integrated with coordinate descent to escape

weak local minima and obtain high-quality sparse solutions. Our specialized C++ implementation is comparable in speed to state-of-the art sparse regression packages such as glmnet.

E0516: **Optimal classification and regression trees**
Presenter: **Jack Dunn**, Massachusetts Institute of Technology, United States
Co-authors: Dimitris Bertsimas
Decision tree methods are widely used for classification and regression problems, but one must choose between methods that are interpretable (such as CART) and those with high accuracy (random forests or boosting). We introduce a new approach for constructing Optimal Classification and Regression Trees based on mixed-integer optimization, and develop high-performance heuristics for these problems that offer significant improvements over traditional greedy approaches and run in comparable times. We show in a collection of synthetic and real-world instances that our Optimal Trees significantly improve upon greedy tree methods like CART, giving solutions that are both interpretable and have accuracies comparable to ensemble methods.

E1118: **Computational properties of solution methods for logistic regression through the lens of condition numbers**
Presenter: **Paul Grigas**, University of California, Berkeley, United States
Co-authors: Robert Freund, Rahul Mazumder
The simple probabilistic model underlying logistic regression suggests that it is most natural and sensible to apply logistic regression when the data is not linearly separable. Building on this basic intuition, we introduce a pair of condition numbers that measure the degree of non-separability or separability of a given dataset (in the setting of binary classification). When the sample data is not separable, the degree of non-separability naturally enters the analysis and informs the properties and convergence guarantees of a wide variety of computational methods, including standard first-order methods such as steepest descent, greedy coordinate descent, etc. When the sample data is separable – in which case standard logistic regression will break down – the degree of separability can be used to show, rather surprisingly, that standard first order methods with implicit regularization also deliver approximate-maximum-margin solutions with associated computational guarantees as well. Finally, in order to further enhance our understanding of the computational properties of several methods, we also take advantage of recent new results on self-concordant-like properties of logistic regression due to Bach.

E0564: **Testing degree corrections in stochastic block models**
Presenter: **Rajarshi Mukherjee**, University of California Berkeley, United States
Co-authors: Subhabrata Sen
Thresholds for degree corrections in Stochastic Block Models are studied in the context of a goodness of fit problem. When degree corrections are relatively dense, a simple test based on the total number of edges is asymptotically optimal. For sparse degree corrections in non-dense graphs, simple degree based Higher Criticism Test is optimal with sharp constants. In contrast, for dense graphs, the optimal procedure runs in two stages. It involves running a suitable community recovery algorithm in stage 1, followed by a Higher Criticism Test based on a linear combination of within and across (estimated) community degrees in stage 2. The necessity of the two-step procedure is demonstrated by the failure of the ordinary Maximum Degree Test in achieving sharp constants. As necessary tools, we also derive asymptotic distribution of the Maximum Degree in Stochastic Block Models along with moderate deviation and local central limit type asymptotics of positive linear combinations of independent Binomial random variables.

EO228 Room CLO 101 NOVEL BAYESIAN APPLICATIONS AND METHODS Chair: Christopher Hans

E0473: **Spatio-temporal modeling of urban data: A case study in Philadelphia**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States
Throughout history there have been many perspectives on the approach to planning cities, with a notable clash between dense, organically-formed urban spaces and suburban space planned around the automobile. Urban data analysis has been recently improved through publicly available high resolution data, allowing us to empirically investigate urban design principles of the past half-century. We are currently focussed on the analysis of local neighborhood features, including crime, land use, zoning, business and population demographics. We will discuss one particular direction: spatial modeling of the change in safety over the past decade. The city of Philadelphia is an interesting case study, with its rapid urban development in the last decade.

E0450: **Bayesian hierarchical modelling of sparse count processes with applications in retail analytics**
Presenter: **Ioanna Manolopoulou**, University College London, United Kingdom
Co-authors: Gordon Ross, James Pitkin
In retail analytics, slow-moving-inventory (SMI) refers to goods which rarely sell, resulting in very sparse count processes. Forecasting the sales of such goods is challenging, because traditional predictive models rely on large enough sales volumes to be accurate. We develop modelling, inferential and predictive methods able to learn the dynamics of sparse count processes for SMI products with few to no sales. We flexibly introduce covariates into a previous self-exciting model for sparse processes. We extend the model to include a cross-excitation contribution that allows differing series to excite one another, capturing the process of intertwined contemporaneous excitation dynamics. We integrate individual products into a Bayesian hierarchical model that accommodates shrinkage and information passing across differing sparse count process, without requiring the data for each product to exist over the same time period. We illustrate our methods on a retail analytics dataset from a major supermarket chain in the UK.

E0487: **Bayesian regression tree models for causal inference**
Presenter: **Carlos Carvalho**, The University of Texas at Austin, United States
Co-authors: Richard Hahn, Jared Murray
A semi-parametric Bayesian regression model is developed for estimating heterogeneous treatment effects from observational data. Standard nonlinear regression models, which may work quite well for prediction, can yield badly biased estimates of treatment effects when fit to data with strong confounding. Our Bayesian causal forests model avoids this problem by directly incorporating an estimate of the propensity function in the specification of the response model, implicitly inducing a covariate-dependent prior on the regression function. This new parametrization also allows treatment heterogeneity to be regularized separately from the prognostic effect of control variables, making it possible to informatively shrink to homogeneity,in contrast to existing Bayesian non- and semi-parametric approaches.

E1017: **Partial identification of causal effects in grouped data with unobserved confounding**
Presenter: **Jennifer Hill**, New York University, United States
The unbiased estimation of a treatment effect in the context of observational studies with grouped data is considered. When analyzing such data, researchers typically include as many predictors as possible, in an attempt to satisfy ignorability, and so-called fixed effects (indicators for groups) to capture unobserved between-group variation. However, depending on the mathematical properties of the data generating process, adding such predictors can actually increase treatment effect bias if ignorability is not satisfied. Exploiting information contained in multilevel model

estimates, we generate bounds on the comparative potential bias of competing methods, which can inform model selection. Our approach relies on a parametric model for grouped data and an omitted confounder, establishing a framework for sensitivity analysis. We characterize the strength of the confounding along with bias amplification using easily interpretable parameters and graphical displays. Additionally we provide estimates of the uncertainty in the derived bounds and create a framework for estimating causal effects with partial identification.

E1066: **Hierarchical array priors for ANOVA decompositions of cross-classified data**
Presenter: **Alexander Volfovsky**, Duke University, United States
ANOVA decompositions are a standard method for describing and estimating heterogeneity among the means of a response variable across levels of multiple categorical factors. In such a decomposition, the complete set of main effects and interaction terms can be viewed as a collection of vectors, matrices and arrays that share various index sets defined by the factor levels. For many types of categorical factors, it is plausible that an ANOVA decomposition exhibits some consistency across orders of effects, in that the levels of a factor that have similar main-effect coefficients may also have similar coefficients in higher-order interaction terms. In such a case, estimation of the higher-order interactions should be improved by borrowing information from the main effects and lower-order interactions. To take advantage of such patterns, a class of hierarchical prior distributions for collections of interaction arrays is introduced which can adapt to the presence of such interactions. These prior distributions are based on a type of array-variate normal distribution, for which a covariance matrix for each factor is estimated. This prior is able to adapt to potential similarities among the levels of a factor, and incorporate any such information into the estimation of the effects in which the factor appears.

EO423 Room CLO 102 CONTRIBUTIONS TO THE ESTIMATION PROBLEM IN STOCHASTIC SYSTEMS Chair: Raquel Caballero-Aguila

E0746: **Signal filtering over sensor networks with random delays and loss compensations: Distributed and centralized framework**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaan, Spain
Co-authors: Aurora Hermoso-Carazo, Josefa Linares-Perez
Recently, the estimation problem in sensor networks has received considerable attention due to its great number of application fields. In a network environment, the inaccuracy of the measurement devices and the limitations of the communication resources usually cause random phenomena that impair the performance of the estimators; consequently, many interesting challenges arise in the research of the fusion estimation problem over sensor networks featuring random imperfections. When the measurements are subject to transmission losses, a hot topic is how to compensate them; the more common compensation mechanisms consist of processing either nothing (non-compensation) or the latest successfully transmitted measurement. The aim is to study the distributed and centralized fusion filtering problems from measured outputs with uncertainties modelled by random matrices and correlated noises. At each sampling time, every sensor output is sent just once to a local processor and, due to random transmission failures, one-step delays and, consequently, packet losses may occur; it is assumed that the losses are non-consecutive and compensated with the last measurement received at the local processor. Modelling the random delays and losses by different sequences of Bernoulli random variables, the estimation is addressed by a covariance-based approach. In a simulation example, the results are compared with those obtained when the packet losses are not compensated.

E1352: **Least-squares quadratic filtering in linear stochastic systems with random parameter matrices and correlated noises**
Presenter: **Irene Garcia-Garrido**, Universidad de Jaan, Spain
Co-authors: Raquel Caballero-Aguila, Josefa Linares-Perez
A quadratic filtering algorithm in linear discrete-time stochastic systems with random parameter matrices is designed. The additive noises involved in the system are supposed to be autocorrelated and cross-correlated. To address the quadratic estimation problem under the least-squares (LS) optimality criterion, the system hypotheses must include the knowledge of the fourth-order moments of the processes involved. Defining a suitable augmented system by stacking the original state and observation vectors with their second-order Kronecker powers, the quadratic estimation problem is then reformulated as a linear estimation problem, whose approach requires to study the second-order statistical properties of the processes involved in this augmented system. By an innovation approach, a recursive algorithm for the LS linear filter of the augmented state based on the augmented observations is derived, from which the required LS quadratic filter of the original state is obtained. The proposed algorithm can be applied to multisensor systems with random uncertainties in both the state and measurement equations. The performance of the proposed estimator is illustrated by a numerical simulation example where a scalar state process is estimated from multisensor missing measurements. The quadratic estimation accuracy is analyzed in terms of their error variances for different missing probabilities and the superiority of the quadratic filter in comparison with the linear one is also shown.

E1562: **A filtering algorithm for systems with random transmission delays modeled by multi-state Markov chains**
Presenter: **Maria Jesus Garcia-Ligero**, Universidad de Granada, Spain
Co-authors: Aurora Hermoso-Carazo, Josefa Linares-Perez
An unavoidable problem in communication networks is the existence of delays in the arrival measurements. The delay may be deterministic or random, although in the most practical cases, such as mobile communications, exploration seismology, between others, the delay is random, being modeled by a stochastic process. Traditionally, the randomly delayed measurements have been modeled by independent random variables. However, in real communication systems, current time delays are usually correlated with the previous ones; a reasonable way to model the dependence on the delays is to consider them as homogeneous Markov chains. In this context, signal estimation algorithms have been derived considering that the measurements can be delayed by one, two or more sampling times. We generalize this situation to the case of measurements that can be delayed by one, two or more sampling times. Specifically, the least-squares linear estimation of signals is addressed considering that the delays are modeled by homogeneous multi-state Markov chains. The linear filter of the signal is derived by using the information provided by the covariance functions of the process involved in the observations, as well as the probability distribution of Markov chain modeling the delays.

E1769: **Spatial point processes estimation in functional spaces**
Presenter: **Maria Pilar Frias Bustamante**, University of Granada, Spain
Co-authors: Maria Dolores Ruiz-Medina, Antoni Torres
Previous results are extended concerning Cox processes in functional spaces and minimum contrast estimation of spatial processes from tapered data. Specifically, Log-Gaussian spatial point processes with SARH(1) intensity are introduced. Minimum-contrast componentwise estimation of the log-Gaussian SARH(1) intensity is achieved. Spatial functional prediction of the corresponding functional Cox-process is then performed. A simulation study is undertaken to investigate the asymptotic properties of the proposed spatial functional predictor.

E0969: **Distributed estimation in networked systems with random parameter matrices and transmission packet dropouts**
Presenter: **Aurora Hermoso-Carazo**, Universidad de Granada, Spain
Co-authors: Raquel Caballero-Aguila, Josefa Linares-Perez, Zidong Wang
Over the last few decades, the estimation problem in networked systems has attracted increasing attention and different kinds of fusion estimation algorithms have been reported. Recently, one of the most challenging and fertile research fields is the distributed estimation problem in sensor networks, where the sensor nodes are spatially distributed according to a given network topology, so every sensor is only connected and exchange

information with its neighbors. In this framework, the aim concerns the distributed filtering problem in discrete-time stochastic systems over a sensor network with a given topology, assuming that the measured outputs are perturbed by random parameter matrices and correlated additive noises and, moreover, random packet dropouts may occur during the data transmission among the sensor nodes through the different network communication channels. By an innovation approach, using the last measurement received if a data packet is lost, an intermediate least-squares filter is designed at each sensor node using its own measurements and those from its neighboring nodes. After that, every sensor collects the intermediate filters that are successfully received from its neighbors and the final distributed filter is obtained as the least-squares matrix-weighted linear combination of the intermediate ones within its neighborhood.

EO372 Room CLO 203 ORDINAL DATA MODELING: PERSPECTIVES AND APPLICATIONS Chair: Rosaria Simone

E0329: **Joint model-based clustering for ordinal survey data**
Presenter: **Bettina Gruen**, Johannes Kepler University Linz, Austria
Co-authors: Sara Dolnicar
Survey data collected is typically ordinal in nature. As such, it is susceptible to response styles. Response styles are consistent tendencies displayed by survey respondents when responding which are not related to the specific item content. When clustering ordinal survey data, ignoring response styles can lead to clusters which do not differ in beliefs, but merely in how cluster members use survey answer options and which possibly occur in addition to the belief clusters. We propose a finite mixture model which simultaneously clusters and corrects for response styles, permits heterogeneity in both beliefs and response styles, accommodates a range of different response styles, does not impose a certain relationship between the response style and belief clusters, and is suitable for ordinal data. The performance of the model is tested using both artificial and empirical survey data.

E0968: **Measures of ordinal association in two-way contingency tables**
Presenter: **Maria Kateri**, RWTH Aachen University, Germany
For IxJ contingency tables with ordinal classification variables, measures of association are discussed that are based on generalized odds ratios (GORs). The GORs are linked to association models and expressed in terms of their parameters. Since the maximum likelihood estimation of the association models' parameters requires the use of iterative procedures, closed-form approximations to the maximum likelihood estimators (MLEs) are considered that allow for a handy estimation of the corresponding association measures. In the literature there exist some measures of this type that are related to the uniform association (U) model. We introduce new measures, based on the row effect association (R) model, that are more flexible in capturing structures of underlying association other than the uniform. Furthermore, we propose alternative closed-form estimators for measures based on U and R models. The new measures are illustrated via examples while the various closed-form approximations are compared to the MLEs via extensive simulation studies.

E0887: **A review of panel data models with a Markov dependent structure for univariate and multivariate ordinal responses**
Presenter: **Fulvia Pennoni**, University of Milano-Bicocca, Italy
The purpose is to review the theory related to the Markov chain model and to the latent Markov model with univariate or multivariate ordinal responses. It frequently happens on social and medical surveys, and it is also common to encounter such data on other fields of applications by which the order can be made according to subject matter knowledge. We review how to assess the influence of covariates according to a chosen link function based on local, global, or continuation logits, or by an ordered probit link function. In this way, the model parameters can retain the ordinal structure imposed by the data. We show the statistical algorithms for the parameters estimation and through examples based on real data we show how predictions can be achieved.

E0900: **A unifying perspective to model preference and evaluation data**
Presenter: **Domenico Piccolo**, University of Naples Federico II, Italy
Co-authors: Stefania Capecchi, Maria Iannario, Rosaria Simone
"All models are wrong, but some are useful": this statement made by G.E.P. Box is a general benchmark that should be supposed in testing different models. Thus, it should be assumed to drive the model specification step. More specifically, the motivation comes from the idea of disclosing the advantageous traits of different modelling strategies for ordinal variables arising in survey analysis of preference and evaluation data. The aim is to compare two paradigms generated by different perspectives: the consolidated and multifold class of cumulative models and the flexible and parsimonious approach afforded by CUB mixture distributions, whose core is to account for uncertainty and heterogeneity. A simulation study and empirical evidence are discussed in order to check for behaviours, strengths, shortcomings and performances of the competing modelling techniques. In conclusion, a unifying proposal is advanced.

E1372: **Modelling of dispersion and response styles in ordinal regression**
Presenter: **Gerhard Tutz**, LMU, Germany
In ordinal regression the focus is typically on location effects, potential variation in the distribution of the probability mass over response categories referring to stronger or weaker concentration in the middle is mostly ignored. If dispersion effects are present but ignored goodness-of-fit suffers and, more severely, biased estimates of location effects are to be expected. A model specification is proposed that explicitly links varying dispersion to explanatory variables. It is able to explain why frequently some variables are found to have category-specific effects if the cumulative type model is used. For repeated measurements, which are used in survey-based research, dispersion effects may be seen as response styles that represent specific answering patterns of respondents. We consider an extension of the Partial Credit Model that explicitly accounts for response styles. A common problem in partial proportional odds models is the selection of the effect type, each covariate can be equipped with either a simple, global effect or a more flexible and complex effect which is specific to the response categories. A general penalty framework is proposed that allows for an automatic, data-driven selection between global and category-specific effects in all types of ordinal regression models.

EO651 Room MAL 414 STATISTICAL METHODS FOR MOBILE HEALTH APPLICATIONS Chair: Jaroslaw Harezlak

E0870: **Challenges and opportunities in the analysis of continuous glucose monitoring data**
Presenter: **Irina Gaynanova**, Texas A and M University, United States
Diabetes affects approximately 415 million people worldwide, and can increase the likelihood of long-term complications such as cardiovascular disease and nerve damage. The control of blood sugar levels is crucial in reducing these risks, and is the primary focus of diabetes management. Traditionally, glucose levels are measured only 1-4 times a day, typically around meal times, by means of a finger prick. This strategy however does not provide patients adequate information to correct and control fluctuations throughout the day. Continuous glucose monitors (CGM) can provide regular glucose level readings as often as every 5 minutes and therefore have great potential in improving diabetes management. Nevertheless, data complexity and measurement errors present significant difficulties in fully utilizing CGM data for clinical decision making. We will present statistical methods for assessing glucose control from CGM data as well as highlight the challenges associated with the corresponding analysis.

E1130: **A functional data analysis framework for objectively measured physical activity data from accelerometry**
Presenter: **Chongzhi Di**, Fred Hutchinson Cancer Research Center, United States

In large-scale epidemiological studies, it is increasing common to record physical activity objectively by wearable accelerometers. Accelerometry data are time series that allow more precise measurement of the intensity, frequency and duration of physical activity than self-reported questionnaires. However, standard analysis often reduce the high-resolution data into a few simple summary measures, which depends on choices of cut points and can be oversimplied. We develop a functional data framework for the analysis of accelerometry data. We first introduce functional indices to describe the profile of activity intensity, frequency and duration. These indices are then used as outcomes or predictors in functional regression analysis, which allows estimation of detailed dose-response relationship between activity patterns and health outcomes. These methods are motivated by and applied to the Objective Physical Activity and Cardiovascular Health Study among older women, where the aim is to study the association between objectively measured physical activity and cardiovascular diseases.

E0944: **Classification of human activity via spherical representation of accelerometry signal**
Presenter: **Michal Kos**, University of Wroclaw, Poland
Co-authors: Jaroslaw Harezlak, Malgorzata Bogdan, Nancy Glynn

Human health is strongly associated with person's lifestyle and level of physical activity. Therefore, characterization of daily human activity is an important task. The accelerometer is a wearable device which enables precise measurements of the acceleration changes over time of a body part to which it is attached. It can collect over a 1,000,000 observations per hour. The signal from an accelerometer can be used to classify different types of activity. We propose a novel procedure of activity classification, which is based on the spherical representation of the raw accelerometry data. Accurate classification information is obtained from the angular part of a signal, which is partially summarized via the spherical variance. One of the method's main properties is its ability to provide classification of short term activities. The classification accuracy of our method is 90% for the within-subject level and 83% for the between-subject level. In summary, the major contribution to the accurate classification of different physical activity kinds is the use of the spherical variance exhibiting rotational invariance. This property makes it insensitive to the shifts of an accelerometer location.

E0994: **Walking recognition via continuous wavelet transforms applied to the longitudinal intervention study**
Presenter: **Marcin Straczkiewicz**, School of Public Health-Bloomington, Indiana University, United States
Co-authors: Christopher Sorensen, Jacek Urbanek, Beau Ances, Ciprian Crainiceanu, Jaroslaw Harezlak

Recent advancements in accelerometry provide researchers with more objective ways to quantify physical activity (PA) than the commonly used questionnaires. Wearable devices enable the quantification of the behavior in a free-living environment. However, the common summary measures, e.g. activity counts, lose crucial information from the signal. We developed an algorithm (RoWW - Recognition of Walking using Wavelets) based on the Continuous Wavelet Transform (CWT) to classify the walking vs. non-walking periods using raw accelerometry data. Our algorithm quantifies the characteristics of the walking periods including its periodicity, duration, speed and intensity. Our CWT-based algorithm is sensitive to subtle changes in periodic content of a signal and detection of its breakdown points or signal discontinuities. RoWW automatically fragments the time-frequency display of a given signal and identifies the time segments with the pronounced periodicity. We apply RoWW to study the influence of two types of behavioral interventions on walking habits among a cohort of HIV-infected patients, whose data were collected for one week at baseline and at a 3-month visit.

E1050: **Algorithm for differentiation of walking and stair climbing based on the raw accelerometry data**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: William Fadel, Jacek Urbanek, Xiaochun Li, Steven Albertson

Wearable accelerometers provide an objective measure of human physical activity. They record high frequency unlabeled three-dimensional time series data. We extract meaningful features from the raw accelerometry data and based on them develop and evaluate a classification method for the detection of walking and its subclasses, i.e. level walking, descending stairs and ascending stairs. Our methodology is tested on a sample of 32 middle-aged subjects for whom we extracted features based on the Fourier and wavelet transforms. We build subject-specific and group-level classification models utilizing a tree-based methodology. We evaluate the effects of sensor location and tuning parameters on the classification accuracy of the tree models. In the group-level classification setting, we propose a robust feature inter-subject normalization and evaluate its performance compared to unnormalized data. The classification accuracy on the subject-specific level was on average 87.6%, with the ankle-worn accelerometers showing the best performance with an average accuracy 90.5%. At the group-level, the average classification accuracy using the normalized features was 80.2% compared to 72.3% for the unnormalized features. In summary, a framework is provided for better use and feature extraction from the raw accelerometry data to differentiate among different walking modalities.

EO680 Room MAL 416 MODERNIZING ASSESSMENT AND INSTRUCTION IN STATISTICS Chair: Erin Blankenship

E0267: **Revising the GAISE: Guidelines for Assessment & Instruction in Statistics Education**
Presenter: **Megan Mocko**, University of Florida, United States

In May 2005, the GAISE College Report was endorsed by the American Statistical Association. The report included six recommendations for teaching introductory statistics as well as four appendices. This report has had a deep impact on statistics classrooms and educational research. However much has changed in the discipline since 2005, for example online learning, flipped classrooms as well as Big Data are now quite common. A group of eleven volunteers went about revising the report using feedback from the statistical education committee and recent statistics education literature. Overall, it was decided to keep the original six recommendations (with small changes to wording and order), but to add two additional emphasis: 1.) Teach statistics as an investigative process of problem-solving and decision-making; 2.) Provide students with experience with multivariable thinking. The preexisting appendices were updated, and additional appendices were added to support instructors teaching introductory statistics. The aim is discuss in more detail why it was decided to revise the GAISE College Report, how the document was revised and give more detail on what revisions took place.

E0301: **Reflections on best practices for teaching the algebra-based introductory statistics course**
Presenter: **Nathan Tintle**, Dordt College, United States

The last three decades have seen substantial changes in statistical practice, teaching methods and understanding of how people understand probabilistic and statistical concepts. The intersection of these developments has led to the statistics education community increasingly focusing on the use of simulation-based methods, including bootstrapping and permutation tests, to illustrate core concepts of statistical inference within the context of the overall research process. This new focus presents an opportunity to address documented shortcomings in introductory level statistics courses. We will (1) discuss the motivation and rationale behind the simulation-based approach, (2) share some concrete examples of how the approach works and can be integrated into existing courses, (3) present research evidence of its effectiveness at impacting students conceptual understanding and attitudes post-course and in the months following the courses completion and, (4) share a wealth of instructional resources available to support instructors trying out and using these approaches. These comments will be made within the larger context of other best practices, trends and debates about strategies for teaching the algebra-based introductory statistics course.

E0770: **Cutting through the theory: Emphasizing statistical thinking in mathematical statistics**
Presenter: **Erin Blankenship**, University of Nebraska-Lincoln, United States
Co-authors: Jennifer Green
More than 10 years after the release of the original Guidelines for the Assessment and Instruction in Statistics Education (GAISE) College Report, the proposed reforms have been widely embraced by the statistics education community for introductory statistics courses. Although the report is geared toward the introductory course, the GAISE recommendations are just as relevant for higher level statistics courses. We discuss implementing the reforms in the two-semester calculus-based undergraduate mathematical statistics sequence. While all of the GAISE recommendations are applicable, we focus on methods for incorporating conceptual understanding and active learning into these courses, which are typically taught in a traditional lecture style. Our course revision does not require substantial changes in course content, so the necessary theoretical content is not sacrificed. We will describe activities and teaching strategies we have used to promote student understanding and problem solving, as well as share student reflections on the course revisions.

E0721: **Educating educators: Developing teachers of statistical thinking**
Presenter: **Jennifer Green**, Montana State University, United States
Co-authors: Erin Blankenship
The Guidelines for the Assessment and Instruction in Statistics Education (GAISE) PreK-12 Report has helped to shape changing standards and curricula for school-level (i.e, pre-university) statistics. The guidelines emphasize the role of variability in the statistical problem-solving process of formulating questions, collecting and analyzing data, and interpreting results. With a focus on the use real data and conceptual understanding, these guidelines and standards represent a critical shift from using formulaic approaches to using data-oriented approaches to teach modern school-level statistics. However, many teachers have not received the training needed to implement such an approach, creating demands to develop and provide authentic opportunities for teachers to engage in the statistical problem-solving process and develop their own statistical thinking and literacy skills. We will discuss the GAISE recommendations for teaching school-level statistics and their implications for preparing teachers of modern statistics curricula. We will also briefly describe the courses we developed to support teachers of statistics, and share reflections and insights from our research on teachers' preparation and development.

E1004: **Teaching statistical thinking to statistics graduate students**
Presenter: **Walter Stroup**, University of Nebraska, United States
Most statistics graduate programs have a core curriculum that includes, at a minimum, theory (mathematical statistics and probability) and statistical methods (linear models and - possibly - design of research studies). George Box once commented that at some point, the discipline of statistics had been hijacked by mathematics. Although our discipline has made efforts to address Professor Box's concern, it is still true that students earning Masters degrees in statistics tend to know far more about the mathematics and computing of statistics than they do about what I call "the statistics of statistics". Ideally, statistical methods courses should be viewed as science courses. Statistics is not mathematics, nor is it a "branch of mathematics", as many dictionaries misleadingly define it. Statistics is a bridge between mathematics and science (as broadly defined). In developing new courses for the University of Nebraska's revision of its Statistics MS curriculum, we have developed teaching tools and student activities designed to help students "get inside the scientist's head". Students learn the methods and their supporting theory, but more importantly, they learn how to use these methods to help scientists answer the questions that motivate them to plan studies and collect data.

EO302 Room MAL 421 ROBUSTNESS IN DATA SCIENCE Chair: Valentin Todorov

E0390: **Refactoring the FORTRAN code for LTS and MCD Algorithm in R**
Presenter: **Peter Ruckdeschel**, University of Oldenburg, Germany
Co-authors: Valentin Todorov
Recent progress in a complete refactoring of the (fast) LTS and MCD Algorithm completely into R is reported. For iid data, robust covariances and robust regression methods are readily available in many statistical software packages to amend the instability of the classical procedures in the presence of outliers. In R, this is true within the robustbase and rrcov packages. In both packages, code interfaces to original FORTRAN code. When it comes to weighted data as it arises, e.g., in stratified sampling, in model based clustering, Gaussian mixture models, or Hidden Markov Models, the Minimum covariance determinant (MCD) estimator as one of the most important robust alternatives so far is not available. The same is true for the LTS estimator for regression in a weighted data setting. We close this gap, (re-)implementing the (fast)MCD and (fast)LTS estimators from FORTRAN into pure R, with additional coverage of weighted data. Moreover the new pure R code lends itself more easily for future extensions like more general data structures for spares matrices than the current FORTRAN code. As with the reference MCD and LTS code in R/FORTRAN, we provide a reweighting step to achieve higher efficiency in the ideal model than the raw MCD / LTS estimator.

E1315: **Finding conics in noisy images**
Presenter: **Agustin Mayo-Iscar**, Universidad de Valladolid, Spain
Co-authors: Luis Angel Garcia-Escudero, Alfonso Gordaliza
Trimming is a useful tool for robustifying different statistical methodologies. For the joint location-scatter estimation and the linear regression model estimation, MCD and LTS are well known trimming proposals. Now, we are interested in applying trimming for finding conics in noisy images. In order to do it, we have designed an algorithms, for different situations based on the robustification of the corresponding procrustes approaches. As usual, when applying this technique, the level of trimming will be an input parameter which should be provided in advance. We have got empirical evidences about how the methodology works obtained after applying these procedures both to artificial data and real data.

E1338: **fsdaR: making the FSDA toolbox available to R users**
Presenter: **Emmanuele Sordini**, Joint Research Centre of the European Commission (JRC), Italy
Co-authors: Valentin Todorov, Aldo Corbellini
The FSDA (Flexible Statistics for Data Analysis) toolbox is a software library that extends MATLAB and its Statistics Toolbox to support a robust and efficient analysis of complex datasets, affected by different sources of heterogeneity. FSDA was born around the forward search approach, and has evolved into a comprehensive and computationally efficient software package for robust statistics. FSDA provides tools for regression and multivariate analysis, robust transformations, cluster analysis and model selection. FSDA provides a rich set of graphical features not available in R, such as dynamic brushing, linking, etc., particularly useful for exploratory data analysis. The only downside is that FSDA is based on commercial software (MATLAB), which is not so appealing to the majority of the statistical community, where R is more widespread. Therefore, in 2016 we developed a prototype in which we demonstrated the feasibility of calling MATLAB functions in R in order to make selected features of FSDA available to R developers. The first prototype has since evolved into a package (named fsdaR) compliant with R coding and calling standards (object-oriented approach, formula notation, etc.), providing to R users selected features of the FSDA toolbox along with some of its advanced graphical capabilities. We will focus on i) The structure of the fsdaR package; ii) computational and graphical features of FSDA available in R through fsdaR, iii) A live demo.

E1490: **Real-time DetMCD-based classification of NIR-spectra**
Presenter: **Iwein Vranckx**, KU Leuven, Belgium
Co-authors: Mia Hubert, Bart de Ketelaere
Modern, state-of-the-art sorting machines process several gigabytes of chemometric measurements in milliseconds, frequently pushing the boundaries of today's available computational power. This is further complicated by the fact that classic machine classifiers are typically not outlier-resistant. Given that they are practically omnipresent in industrial datasets, it is clear that industrial classifiers are often operating at sub-optimal efficiency.The goal is the equipping of food sorting machines with robust, DetMCD-based classifiers. In order for this to work, two major improvements are made: Firstly: DetMCD is pretty fast, but the current implementation still requires too much time for online applications. We propose several algorithmic modifications, among which techniques for parallel computing, such that our industrial time constraints are met. Secondly, the outlier detection capability of DetMCD is directly linked with industrial machine precision and should, therefore, be as high as possible. We improve this by using adaptive C-steps. To illustrate the performance of our improved DetMCD estimator, the algorithm is applied to industrial chemometric datasets of various food related products. We demonstrate the corresponding machine efficiency improvements and highlight the improved classification training times.

E1604: **Robust SUR estimation under an independent contamination model**
Presenter: **Fatemah Alqallaf**, Kuwait University, Kuwait
The Seemingly Unrelated Regressions (SUR) model is a wide used estimation procedure in econometrics, insurance and finance, where very often, the regression model contains more than one equation. Unknown parameters, regression coefficients and covariances among the errors terms, are estimate using algorithms based on Generalized Least Squares or Maximum Likelihood, and the method, as a whole, is very sensitive to outliers. To overcome this problem M-estimators and S-estimators are proposed in the literature together with fast algorithms. However, these procedures are only able to cope with row-wise outliers contamination in the error terms, while their performance becomes very poor in the presence of cell-wise contamination and as the number of equations is large. A new robust approach is proposed which is able to perform well under both contamination types as well as is fast to compute. Illustrations based on Monte Carlo simulations and real examples are provided.

EO678 Room MAL 532 SPATIAL POINT PROCESS MODELS AND THEIR STATISTICAL INFERENCE Chair: Dominic Schuhmacher

E1414: **Point patterns in space and space-time: Linear models and change of support**
Presenter: **Jorge Mateu**, University Jaume I, Spain
Some developments are presented in the theory and applications of point patterns both in space and space-time, focussing in particular in the statistical analysis of linear models for point patterns. We first present some methods for analysing individual effects and their interactions of one or several factors/groups on a set of replicated spatial (possibly marked) and spatio-temporal point patterns. We provide some test statistics based on both the K-function and the mark-weighted K-function. Since the distributions of our statistics are analytically intractable, we also develop some non-parametric mechanisms for statistical inference. We then present some connections between a spatio-temporal point process and a spatial marked point process by considering a temporal marked point process and a spatial marked one by setting times and locations as marks, respectively. These marked point processes can be explored through several mark summary functions that provide valuable information about the behaviour of the marks and can be used as exploratory quantities that measure the strength and range of interactions between marks. Finally, we have a look at point patterns defined over linear networks opening new avenues of research in this new type of support. We develop new estimators of the first-order intensity function based on both a kernel approach and a Voronoi-based estimator. Applications to real data will be presented.

E1484: **Treating ecological data structures as thinned point processes**
Presenter: **Janine Illian**, University of St Andrews, United Kingdom
In statistical ecology specific data structures resulting from common survey methods, motivate statistical methodology and software development. The specific survey method as well as the statistical methodology adapt to the practicality of data collection but do not directly reflect the underlying ecological process of interest. This results in highly specialised modelling approaches and little exchange among developers of different strands of methodology. Thinking in terms of the processes that we would like to model rather than in terms of the survey method can yield a flexible class of models. Specifically, ecological processes of interest are structures formed by individuals in space and time, reflecting interaction among individuals and with the environment. Classical point process methodology assumes the entire spatial area of interest has been surveyed and that individuals have been detected uniformly in space. Treating ecological data structures as a thinning of a spatial point process allows us to use methodology developed for spatial point processes. As a result, general and computationally efficient model fitting software such as R-INLA, based on computationally efficient integrated nested Laplace approximation become relevant. The package inlabru has been designed to fit thinned point process models, based on R-INLA. It accommodates a wide range of ecological data structures and has also contributed to widening the class of models that may be fitted with R-INLA

E1428: **Global envelope tests, with emphasis on spatial point processes**
Presenter: **Mari Myllymaki**, Natural Resources Institute Finland (Luke), Finland
Co-authors: Tomas Mrkvicka, Pavel Grabarnik, Ute Hahn
Envelope tests are a popular tool in spatial statistics, where they are used in goodness-of-fit testing. These tests graphically compare an empirical function $T(r)$ with its simulated counterparts from the null model. However, the type I error probability α is conventionally controlled for a fixed distance r only, whereas the functions are inspected on an interval of distances I. We propose global envelope tests on I. These tests allow the a priori selection of the global α and they yield p-values. We further propose the global rank envelope test as a solution to the multiple testing problem for Monte Carlo tests. Therefore the rank test can be used also, for example, for goodness-of-fit tests with several test functions and for groups of point patterns. Furthermore, a new functional ANOVA test, with a graphical interpretation of the result, is obtained as an extension of the rank test. The tests are developed for summary functions of spatial point processes in the first place, but they can in principle be applied to any functional data or curves. An R library GET is provided for performing the tests.

E1446: **Overview of anisotropic point pattern statistics**
Presenter: **Tuomas Rajala**, University College London, United Kingdom
Spatial point pattern data describe the locations of events in space. Analysis of such data usually relies on the assumptions of stationarity and isotropy of the underlying random mechanism. Non-stationary extensions of basic statistical tools have been under active development for a few decades now, but anisotropic tools are scarce. We will take look at what is currently available.

E1469: **Point processes on the sphere**
Presenter: **Ege Rubak**, Aalborg University, Denmark
Point patterns on a global scale require us to adapt both statistical tools, models and software to the non-Euclidean setting on the sphere. We will describe both functional summary statistics such as the analogues of Ripley's K-function, the F-function, the G-function, and the J-function as well as various statistical models on the sphere. Such models include both the reference class of (inhomogeneous) Poisson processes and the interesting determinantal point processes. An efficient software implementation for working with point data on the sphere is shown.

EO136 Room MAL 538 RECENT DEVELOPMENTS IN LATENT VARIABLE MODELS Chair: Giuliano Galimberti

E0405: **Sampling of pairs in composite likelihood estimation for latent variable models for categorical responses**
Presenter: **Irini Moustaki**, London School of Economics, United Kingdom
Co-authors: Ioulia Papageorgiou
Pairwise likelihood estimation has been recently developed for estimating the parameters of latent variable and structural equation models. Pairwise likelihood is a special case of composite likelihood methods that use lower order conditional or marginal log likelihoods. The composite likelihood to be maximised is a weighted sum of marginal or conditional loglikelihoods. Weights can be chosen to be equal or unequal for increasing efficiency. We approach the problem of weights from a sampling perspective. More specifically, we propose a sampling method for selecting pairs that is based on those pairs that contribute more to the total variance from all pairs. We demonstrate the performance of our methodology using simulated data and a real example.

E0568: **Factor model estimation by composite minimization**
Presenter: **Matteo Farne**, University of Bologna, Italy
The problem of factor model estimation in large dimensions is addressed under the low rank plus sparse assumption. Existing approaches based on PCA like POET estimator fail to catch low rank spaces characterized by non-spiked eigenvalues, as in this case the asymptotic consistency of PCA defaults. UNLOREC, an alternative approach based on the minimization of a low rank plus sparse decomposition problem, is shown to produce the covariance estimate with the least possible dispersed eigenvalues among all the matrices having the same rank of the low rank component and the same support of the sparse component. Consequently, if dimension and sample size are fixed, loadings and factor scores estimated via UNLOREC provide the tightest possible error bound. The result is based on the eigenvalue dispersion lemma. The effectiveness of UNLOREC factor estimates is finally explored in an exhaustive simulation study, which clarifies that the gain of UNLOREC is larger as the latent eigenvalues are less spiked and the sparse component is more sparse.

E1002: **Multilevel modelling with level-2 missing data: The relationship between student ratings and teacher feelings/practices**
Presenter: **Carla Rampichini**, University of Florence, Italy
Co-authors: Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Roberta Varriale, Leonardo Grilli
The relationship between student's evaluation of university courses and several characteristics of the student, the course and the teacher is studied. In particular, we exploit data from a survey among professors of the University of Padua collecting information about teacher feelings and practices. Student ratings are nested into teachers, calling for multilevel modelling. However, due to survey non-response, the information about feelings and practices is not available for about half of the teachers, posing a serious issue of missing data at level 2. Note that a standard analysis based on the available observations would discard the entire set of student-level data for teachers with no response to the survey. The problem of missing values in level 2 covariates has received little attention. We focus on Multiple Imputation (MI) methods. Imputation methods for level 2 covariates should assign a value for each level 2 unit to be shared by all nested level 1 units. To this end, a procedure suggested in the literature is to work on two distinct datasets for level 1 and level 2 units, performing separate imputations within each dataset while using the results from one in the other. We consider several imputation strategies, including methods exploiting level 1 information to improve imputation of level 2 missing values.

E0807: **A multithreaded implementation of the EM algorithm for finite mixture models**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia
Finite mixture distributions provide a flexible tool for modelling heterogeneous data. However, parameter estimation via the EM algorithm can become very time-consuming due to the complicated expressions involved on the E-step that are numerically expensive to evaluate. We develop a block implementation of the EM algorithm that facilitates the calculations on the E- and M-steps to be spread across a larger number of threads. We focus on the fitting of finite mixtures of multivariate skew normal and skew t-distributions, and show that both the E- and M-steps in the EM algorithm can be modified to allow the data to be split into blocks. The approach can be easily implemented for use by multicore and multiprocessor machines. The improvement in time performance is illustrated on some real data sets.

EC700 Room MAL B18 CONTRIBUTIONS TO FUNCTIONAL DATA ANALYSIS Chair: Enea Bongiorno

E1711: **Family-wise error rate on domain subsets: A unified framework for local inference in functional data analysis**
Presenter: **Konrad Abramowicz**, Umea University, Sweden
Co-authors: Alessia Pini, Lina Schelin, Simone Vantini, Sara Sjostedt-de Luna
A functional test of the hypotheses H0 against H1 (e.g., a test on parameters of a functional-on-scalar linear model) is considered where the aim is to select the parts of the domain where H0 is violated, while controlling the probability of false discoveries. It is straightforward to define an unadjusted p-value function, associating a p-value with each point of the domain. Such a function only point-wise controls the probability of type I error, so it cannot be used for domain-selection purposes, since it would not provide any global control of the probability of type-I error. That is why the definition of an adjusted p-value function provided with a stronger control is often required. We require the control of the probability of falsely rejecting the null hypothesis on subsets of the domain (control of the family-wise error rate, FWER on subsets). We compare different methods to define the adjusted p-value function. The methods that we discuss belong to a general set of methods based on the following steps: a family S of subsets of the domain is defined; the restriction of the null hypothesis is tested on every element of the family; the adjusted p-value of each point is computed as the maximum p-value of the tests of every element containing that point. We consider several methods where the choice of S is either fixed or data-driven.

E1673: **Classification of functional fragments by regularized linear classifiers with domain selection**
Presenter: **Marco Stefanucci**, University of Rome - Sapienza, Italy
Co-authors: David Kraus
The purpose is to classify functional data into two groups by linear classifiers based on one-dimensional projections of functions. We reformulate the task to find the best classifier as an optimization problem and solve it by regularization techniques, namely the conjugate gradient method with early stopping, the principal component method and the ridge method. We study the empirical version with finite training samples consisting of incomplete functions observed on different subsets of the domain and show that the optimal, possibly zero,misclassification probability can be achieved in the limit along a possibly non-convergent empirical regularization path. Being able to work with fragmentary training data we propose a domain extension and selection procedure that finds the best domain beyond the common observation domain of all curves. In a simulation study we compare the different regularization methods and investigate the performance of domain selection. Our methodology is illustrated on a medical data set, where we observe a substantial improvement of classification accuracy due to domain extension.

E0200: **Hypergeometric-type bootstrap quasi-likelihood for functional longitudinal data: Inference and applications**
Presenter: **Guangbao Guo**, Shandong University, China
Quasi-likelihood to model functional data is utilized in longitudinal settings, which is a challenging problem owing to data sparsity, time irregularity, and infinite dimensionality. We develop a novel bootstrap quasi-likelihood to derive an effective estimator. In particular, we provide a choice of

sampling distribution to optimize the estimated results. Parallel bootstrap and quasi-likelihood approaches are combined to deal with irregularly and sparsely observed functional data. The proposed approach achieves some asymptotic properties under several mild conditions. The errors of the approach are smaller than those of existing approaches. Several simulation studies are conducted to illustrate the approach in the setting of discrete and finite time, using several statistical inference indicators. Finally, the excellent performance of the approach is demonstrated by analyzing real functional longitudinal data. It is shown that this approach could become popular through sampling design and parallel implementation.

E1366: **A Riemannian framework for curves with velocity information: Application to detection of bad runway conditions**
Presenter: **Florence Nicol**, ENAC, France
Co-authors: Stephane Puechmorel
Curves can be made amenable to functional data statistics by representing them as points on the so-called shape manifold, that is formally defined as a quotient of the manifold of immersions. It is not a Hilbert space, but it can be provided with Riemannian metrics that allow geodesic distance computation, although some care must be taken in order to avoid degeneracy of the metric. Furthermore, the manifold constructed that way is generally not complete, so that geodesics may exist only in a neighborhood of a given curve. The purpose is to introduce a new kind of Riemannian metric that is especially adapted to the study of curves where the velocity is a discriminating feature. In such a case, the original shape space approach cannot be used since the parametrization invariance will wipe out the velocity information. Motivated by a real use case where one wants to assess runway adherence condition by observing only the radar tracks of landing aircraft, a partial parametrization invariance is introduced, yielding a bundle shape space model on which a relevant metric can be defined. The design of the metric was based on the equations of motion and reflects the internal structure of the data. The equation of geodesics will be given, along with a practical computation algorithm based on a shooting method. The performance of the approach for low adherence detection will be assessed on a set of simulated trajectories and compared with competing algorithms.

E1377: **Domain selection for functional data classification**
Presenter: **Nicolas Hernandez**, Universidad Carlos III de Madrid, Spain
Co-authors: Gabriel Martos, Alberto Munoz
A domain selection approach is proposed for classification problems in functional data. Consider two samples of random elements $f_1,\ldots,f_n$ and $g_1,\ldots,g_m$ in $L^2(X)$ generated from the functional stochastic models $f_i(x)=\mu_k(x)+\varepsilon_i(x)$ for $i=1,\ldots,n$ and $g_j(x)=\mu_k(x)+\varepsilon_j(x)$ for $j=1\ldots,m$ respectively and defined on the same domain $X=[0,1]$. The function $\mu_k(x)$ is the mean function for $k=f,g$ and $\varepsilon(x)$ is a random and independent functional error that captures the variability within each class. In this setting, we propose to use a local–inner product parametrized by the vector $\theta=(\theta_1,\theta_2)$, with $0\leqslant\theta_1<\theta_2\leqslant 1$, such that, $\langle f,g\rangle_\theta=\int_{\theta_1}^{\theta_2} f(x)g(x)dx$. The proposed inner–product induce a local–metric in the space of random elements $L^2(X)$. The optimization of θ is presented as a domain selection technique, where the optimization goal pursue the minimization of the misclassification error rate when classifying samples of random functions.

EC692 Room MAL 152 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS Chair: Keith Knight

E1826: **Parallel strategies for estimating the vector generalized linear model**
Presenter: **Panagiotis Paoullis**, Frederick University and Cyprus University of Technology, Cyprus
Co-authors: Ana Colubi, Erricos John Kontoghiorghes, Panagiotis Paoullis
Strategies for computing the estimators of Vector Generalized Linear Models (VGLMs) are investigated. VGLMs is a class of regression models that are limited only by the assumption that the regression coefficients enter through a set of linear predictors. Examples of models with this structure are related with univariate and multivariate distributions, time series, categorical data analysis, survival analysis, generalized estimating equations, correlated binary data and nonlinear least squares problems to name but a few. The algorithm employed to find the Maximum Likelihood Estimate (MLE) of the VGLM is based on the iteratively reweighted least squares (IRLS) and the Fisher scoring method. Three new methods for computing the IRLS of the VGLM are presented. The first method transforms the VGLM in each iteration to an ordinary linear model and uses the QR decomposition to find the estimate. The other two employ the generalized QR decomposition to compute the MLE of the VGLM which are formulated as iterative generalized linear least-squares problems. Various algorithms for computing the MLE of the VGLM are proposed. The algorithms are based on orthogonal transformations and exploit efficiently the Kronecker structure of the model matrix and the sparse structure of the working weight matrix. Parallel strategies for the numerical estimation of the VGLM are discussed.

E1663: **Optimization approaches for multiple instance classification**
Presenter: **Annabella Astorino**, Consiglio Nazionale delle Ricerche (CNR), Italy
Co-authors: Antonio Fuduli, Giovanni Giallombardo, Giovanna Miglionico
Multiple Instance Learning (MIL) problems are tackled, where the objective is to categorize bags of points (instances). Differently from supervised classification problems, where the label of each point in the training set is known, in a MIL problem only the labels of the bags are known, whereas the labels of the instances are unknown. Examples of application are in image classification, text categorization, drug prediction. We focus on a binary MIL classification problem, where the objective is to discriminate between positive and negative bags by means of an appropriate hyperplane. Moreover, we assume that a bag is negative if all its instances are negative and positive whenever at least one instance is positive, then only the labels of the instances of the positive bags are to be predicted. Different mathematical programming models have been proposed. We analyze two optimization models: a mixed integer nonlinear programming model and a nonconvex nonsmooth unconstrained optimization one. In particular for the first model we propose to use a Lagrangian Relaxation technique, while for the second we construct a DC (Difference of Convex) formulation and we propose to use appropriate DC methods. We also provide some numerical results on benchmark datasets.

E1635: **Distributional trees and forests**
Presenter: **Lisa Schlosser**, Universitaet Innsbruck, Austria
Co-authors: Torsten Hothorn, Achim Zeileis
To obtain probabilistic predictions about a dependent variable based on some set of explanatory variables, a distributional approach is often adopted where the parameter(s) of the distribution are linked to regressors. In many classical models this often only captures the location/expectation of the distribution but over the last decade there has been increasing interest in distributional regression approaches modeling all parameters including location, scale, and shape. Notably, the GAMLSS framework allows to establish generalized additive models using this approach. However, in situations where variable selection is required and/or there are non-smooth dependencies or interactions (especially unknown or of high-order), it is challenging to establish a good GAMLSS. A more natural alternative would be the application of regression trees or random forests but, so far, no general distributional framework is available for these methods. Therefore, the two frameworks are combined here to distributional trees and forests. Applications to real and artificial data illustrate how it can be employed in practice using the R package "disttree".

E1656: **Exploring and improving document-to-vector embeddings**
Presenter: **Erik Mathiesen**, Octavia.ai, United Kingdom
Co-authors: Chan Ford
With the invention of word2vec, glove, doc2vec and similar techniques vector space embeddings and deep learning techniques have become the de-facto standard for modern NLP applications. Despite this there are very few detailed studies of the performance and efficiency of the many embedding techniques available, especially when applied to real-word tasks. Instead the choice of method used for text to vector embedding is often made based on rule-of-thumb or random choice. Setting out to determine what should be the preferred method for our particular application we, along the way, discovered several new techniques as well as a variety of interesting facts about the existing standard methods. Based on this work we will present a detailed review and evaluation of a number of existing document embedding techniques as well as several completely new techniques that emerged from our study. We will show how our novel methods perform in a range of scenarios compared to the commonly used techniques and conclude when these should be utilised.

E0199: **Elemental estimates, influence, and algorithmic leveraging**
Presenter: **Keith Knight**, University of Toronto, Canada
It is well-known that the ordinary least squares estimate can be expressed as a weighted sum of elemental estimates based on subsets of p observations where p is the dimension of parameter vector. These weights can be viewed as a probability distribution on subsets of size p (from n observations) of the predictors. We derive its lower dimensional distributions and define a measure of potential influence for subsets of observations analogous to the diagonal elements of the hat matrix for single observations. This theory is then applied to algorithmic leveraging, which is a method for approximating the ordinary least squares estimates when both n and p are large. This method draws a sample of size $m \ll n$ from the observations where high leverage observations (according to the diagonals of the hat matrix) are sampled with higher probability; we can then estimate the regression parameters using either ordinary (unweighted) or weighted least squares using the sampled observations. In particular, we provide a theoretical justification complementing the empirical evidence that unweighted estimation generally outperforms weighted estimation.

EC698 Room MAL 153 CONTRIBUTIONS TO TIME SERIES ANALYSIS II Chair: James Taylor

E1749: **A Gini-based time series analysis and test for reversibility**
Presenter: **Amit Shelef**, Sapir Academic College, Israel
Co-authors: Edna Schechtman
Time reversibility is a fundamental hypothesis in time series. Gini-based equivalents for time series concepts that enable to construct a Gini-based test for time reversibility under merely first-order moment assumptions are developed. The key idea is that the relationship between two variables using Gini (as measured by Gini autocorrelations and partial autocorrelations) can be measured in two directions, which are not necessarily equal. This implies a built-in capability to discriminate between looking at forward and backward directions in time series. The Gini creates two bi-directional Gini autocorrelations (and partial autocorrelations), looking forward and backward in time, which are not necessarily equal. The difference between them may assist in identifying models with underlying heavy-tailed and non-normal innovations. Gini-based test and Gini-based correlograms, which serve as visual tools to examine departures from the symmetry assumption, are constructed. Simulations are used to illustrate the suggested Gini-based framework and to validate the statistical test. An application to a real data set is presented.

E1753: **Multistep ahead forecasts: Parsimonious approach using varying coefficient models**
Presenter: **Julija Tastu**, Maersk Line, Denmark
There are two common ways for producing multistep ahead predictions: direct and iterative. The direct is based on using a one-step ahead model multiple times: the prediction from the prior time stamp is used as input when generating the following step ahead forecast. The iterative approach is based on building a separate model for each prediction horizon. This recursive strategy is more efficient if the model is correctly specified which is hard to achieve in practice. The direct approach is generally more robust to model misspecifications and allows to incorporate horizon specific predictors. Computational challenges arise with the direct strategy when the number of horizons grow large. Moreover, treating each prediction horizon independently is not an efficient usage of data, as one would naturally expect the influence of predictors to be similar for the neighbouring prediction horizons. Addressing the above we propose to formulate the problem using varying coefficient models. Model coefficients are then described as smooth functions of prediction horizon, allowing to use information from the neighbouring steps when inferring for the specific horizon of interest. Multistep ahead forecasts can then be obtained by estimating parameters at a smaller number of fitting points. The model is applied operationally when forecasting shipping demand from 1 to 91 days ahead. The results are compared to the direct approach.

E1403: **Intervention variables and stochastic trend in state space models**
Presenter: **Elisa Jorge-Gonzalez**, Universidad de La Laguna, Spain
Co-authors: Enrique Gonzalez-Davila, Raquel Martin-Rivero, Domingo Lorenzo-Diaz
Time series are often affected by external phenomena that cause a temporary or permanent change in the expected behavior of the same. There are different ways in which this type of phenomena or interventions influence the development of the series, depending on the duration and the form of the impact caused. The most common interventions are the step effect, a sudden change with permanent duration, and the impulse, a transient point change. These effects are usually incorporated into the models by including intervention variables that improved estimates, regardless of the method of analysis used in the study of the series. The use of state space models in the analysis of structural time series allows, among others, the specification of distributional conditions on the level and trend of the series. The main objective is to analyze the utility of state space models with stochastic trend and slope when there are intervention variables with different degrees of impulse. For this purpose, simulated data will be used, as well as an example of real data of airplane passengers arriving in the Canary Islands.

E1735: **Forecast combining for multivariate probability distributions**
Presenter: **Xiaochun Meng**, Oxford Unuversity, United Kingdom
Co-authors: James Taylor
The evolution of extreme temperature has significant implications for the ecosystem and society. Short-term forecasts of extreme temperature are needed to support decision-making in a variety of organisations. We first generate forecasts of the joint distribution of the daily minimum and maximum temperature using ARMA-GARCH models fitted to daily data. We then compare these predictions with those of a model fitted to hourly data. Rather than selecting between these approaches, we consider the combination of the distributional forecasts produced by each. Combining forecasts has been shown to be beneficial in many different contexts. The combination can be viewed as a synthesis of information, or as a portfolio of forecasts, where the aim is to diversify the forecast risk. Although there have been many methods proposed for combining point forecasts, there is only a small literature on combining distributional forecasts. Very few papers have considered the forecasting of multivariate distributions. We generalise methods previously proposed for combining forecasts of univariate distributions based on quantiles. This framework allows us to vary the combining weights over the outcome space. We use proper scoring functions to optimise the weights, and we evaluate the calibration and sharpness of the resulting forecasts using theoretical analysis, as well as an empirical study based on European temperature data.

E1055: **Dynamic time warping-based fuzzy clustering with spatial information**
Presenter: **Marta Disegna**, Faculty of Management, Bournemouth University, United Kingdom
Co-authors: Pierpaolo Durso, Riccardo Massari
Clustering of spatial multivariate time series finds application in several fields, such as economics, marketing, finance, and medicine. Classification of such complex data requires to take into account both spatial and time dimensions. In particular, one should consider (1) the spatial nature of the units to be clustered, (2) the characteristics of the space of multivariate time trajectories, and (3) the uncertainty related to the assignment of a geographical unit to a given cluster on the basis of the above complex features. Existing spatial-time clustering models can be distinguished into non-spatial time series clustering based on a spatial dissimilarity measure; spatially constrained time series clustering; density-based clustering; model-based clustering. The aim is to discuss a novel spatially constrained multivariate time series clustering for units characterised by different levels of spatial contiguity. In particular, the Fuzzy Partitioning Around Medoids algorithm with Dynamic Time Warping distance and spatial penalization terms is applied to classify multivariate time series. This method has been theoretically presented together with the illustration of the empirical case study related to the identification of touristic agglomerations of cities and towns belonging to the same macro destination and characterised by similar touristic flows over time.

EC696 Room MAL 539 CONTRIBUTIONS IN APPLIED STATISTICS Chair: Anne Gegout-Petit

E1432: **Kriging with continuous and categorical inputs in R**
Presenter: **Dominik Kirchhoff**, Dortmund University of Applied Sciences and Arts, Germany
The implementation of extended Kriging models for mixed continuous and categorical input variables in R is considered. Kriging models (also known as Gaussian Stochastic Process models) are an important tool in metamodel-based optimization as they provide not only fast predictions for expensive processes (e.g., simulations), but also an uncertainty estimator of these predictions. The original Kriging model, however, can only cope with purely continuous input variables. First extensions exist to incorporate also categorical variables. We consider three approaches called Exchangeable Correlation, Multiplicative Correlation, and Unrestrictive Hypersphere-based Correlation, which are different in terms of their flexibility and computational effort. We also implement two new distance-based methods and discuss the advantages and disadvantages of each of the methods.

E1440: **Estimation of proportion in stratified median ranked set sampling**
Presenter: **Nursel Koyuncu**, Hacettepe University, Turkey
The ranked set sampling is a sampling scheme that involves preliminary ranking of the study variable. It can be more efficient than simple random sampling scheme. In the literature the authors have modified ranked set sampling to get more efficient estimates such as median ranked set, extreme ranked set etc. Stratified sampling is also sampling method which used in the heterogeneous population. We have examined the estimation of population proportion in stratified median ranked set sampling theoretically. To compare the superiority of proposed method, we have conducted a simulation study. In the simulation, a real data set is used which determine the air quality in Ankara since June 2016 to May 2017. Using this data set we estimate the air quality under different sampling scheme. The findings show that our proposed method gives more efficient results than classical stratified random sampling.

E1668: **Complexity of United State of America presidential speeches**
Presenter: **Valerio Ficcadenti**, University of Macerata, Italy
Co-authors: Roy Cerqueti
The aim is to explore the rhetoric dynamics of a large collection of United States of America Presidents' speeches. In particular, speeches are viewed as complex systems and are analyzed through rank-size laws, being the words of each speech ranked in terms of their frequencies. At this aim, a best-fit procedure with Zipf or with Zipf-Mandelbrot laws is performed over the 951 talks individually. Thanks to these estimations, it is possible to reach interesting conclusions on how 45 United States Presidents, from April 30, 1789 (Washington) until February 28, 2017 (Trump), have delivered political messages. Our analysis shows some remarkable regularities, not only inside a given speech, but also between different speeches. We discuss the political and linguistics aspects. The building of the dataset itself represents a relevant step of the study. In this respect, by using a web scraping routing on the Miller Center website, a large span of 978 speeches have been downloaded. After a pre-processing phase, the set is reduced to 951; for each one, the words' frequencies are stored and the analysis is performed.

E1732: **Estimating the effectiveness of a digital commerce advertising campaigns using a Geo-experiment**
Presenter: **Iman Al-Hasani**, Durham University, United Kingdom
A Geo-experiment is an approach used to measure the effectiveness of digital advertising campaigns where a region of interest is partitioned into geographical-targeting areas called Geos. The experiment is conducted in two distinct time periods where in the first time period there is no difference in advertising campaign between Geos, whereas during the second time period the campaign for some selected Geos is modified. The aim is to construct an efficient design for targeting the modified advertising campaign. The challenge is to design the campaign in a robust way which permits estimation of the effectiveness of the modified campaign. The issue is related to an absence of covariates related to socio-economic characteristics or other important unknown characteristics that are likely to affect the probability of making purchases. A stochastic simulation platform has been built for studying the effectiveness of advertising campaigns. However, due to the computational resource required, the use of simulation limits the complexity of study which can be carried out. A theoretical framework is developed to study the implications of unobserved covariates for inferences about estimated effects of the campaign. A proxy model is introduced to link the fitted model, without covariates, and the truth which includes unobserved covariates The proxy model makes possible the application of standard results in the literature on maximum likelihood estimation for mis-specified models.

E1846: **Application of copula-based BINAR models in loan modelling**
Presenter: **Andrius Buteikis**, Vilnius University Faculty of Mathematics and Informatics, Lithuania
Co-authors: Remigijus Leipus
A bivariate integer-valued autoregressive process of order 1 (BINAR(1)) with copula-joint innovations is studied. We analyse different parameter estimation methods and compare them via Monte Carlo simulations with emphasis on estimation of the copula dependence parameter. An empirical application on defaulted and non-defaulted loan data is carried out using different combinations of copula functions and marginal distribution functions covering the cases where both marginal distributions are from the same family, as well as the case where they are from different distribution families.

EC704 Room MAL 540 CONTRIBUTIONS TO BIOSTATISTICS Chair: Joyce Niland

E1627: **Mining personalized treatment effects by gradient boosting tree**
Presenter: **Shonosuke Sugasawa**, The Institute of Statistical Mathematics, Japan
Co-authors: Hisashi Noma
It has been recognized that treatment effects of the same treatment are often heterogeneous among patients, and mining the personalized treatment effects is essential for achieving personalized and precision medicine. Here we propose an effective way to estimate the personalized treatment effects by using the gradient boosting tree method known as a powerful nonparametric regression method in machine learning field. We estimate relationships between outcome and individual covariates in treatment and control (alternative) groups, respectively, in randomized clinical trials, and combine them to obtain the personalized treatment effect estimates. We apply the proposed method together with some existing methods to simulated data set and real trial data.

E1416: **Dimensionality and topology of brain-wide attractor dynamics**
Presenter: **Satohiro Tajima**, University of Geneva, Switzerland
Brain-wide, complex neural dynamics are considered crucial for cognitive functions. To understand the principles of human-like high-level cognitions, we need not only to advance recording technologies to track high-dimensional collective activities of neurons, but also to develop new theoretical perspectives to reveal their structures. We present an approach to understanding the structures and functions of dynamics based on data analysis methods combining the delay-embedding theorems in nonlinear dynamics and algebraic topology. The method enables us to characterize the nonlinear couplings among reconstructed attractor dynamics in wide-field electrophysiological data. This reveals a universal hierarchy of cross-area interactions and topological properties in (un)conscious brain dynamics, demonstrating its wide application to deciphering complex biological systems.

E1637: **Box-Cox t random intercept model for estimating usual nutrient intake distributions**
Presenter: **Silvia Ferrari**, University of Sao Paulo, Brazil
Co-authors: Giovana Fumes, Jose Eduardo Corrente
The issue of estimating usual nutrient intake distributions and prevalence of inadequate nutrient intakes is of interest in nutrition studies. Box-Cox transformations coupled with the normal distribution are usually employed for modeling nutrient intake data. When the data present highly asymmetric distribution or include outliers, this approach may lead to implausible estimates. Additionally, it does not allow interpretation of the parameters in terms of characteristics of the original data and requires back transformation of the transformed data to the original scale. An alternative approach is proposed for estimating usual nutrient intake distributions and prevalence of inadequate nutrient intakes through a Box-Cox t model with random intercept. The proposed model is flexible enough for modeling highly asymmetric data even when outliers are present. Unlike the usual approach, the proposed model does not require a transformation of the data. A simulation study suggests that the Box-Cox t model with random intercept estimates the usual intake distribution satisfactorily, and that it should be preferable to the usual approach particularly in cases of highly asymmetric heavy-tailed data. In applications to data sets on intake of 19 micronutrients, the Box-Cox t models provided better fit than its competitors in most of the cases.

E1730: **Change-points in heart rate variability: application in critical care patients**
Presenter: **Ana Paula Rocha**, Univ Porto - Fac Ciencias, Portugal
Co-authors: Margarida Pereira
The detection of change points in time series analysis and signal processing is a relevant problem in various areas such as economy, finance and biomedical applications. Heart Rate Variability (HRV) time series are complex, namely display non-stationary characteristics, exhibit long range dependence in the mean and conditional heteroscedasticity. Previous results suggest that structural breaks detection can be useful in HRV monitoring. Herein are reviewed several methods for segmenting nonstationary HRV time series, considering the case of multiple change point detection. Some approaches are based on the detection and estimation of abrupt changes in mean, root-mean-square level, standard deviation and slope; alternatively, can be used minimum description length principles and time varying spectral tools. The methods are applied in some typical clinical situations from intensive care patients, where segmentation and modelling of HRV may be used as an auxiliary tool in the assessment of the patients evolution and its management.

E1293: **The intra and interclass correlation coefficients for measuring the familial relatedness in genetic studies**
Presenter: **Ozge Karadag**, Hacettepe University, Turkey
Co-authors: Serpil Aktas
Most of the traditional statistical methods have been developed making certain assumptions about the dependency structure of the observations, such as independently identically distributed response variables. However, in many fields of applied sciences, especially in biomedical or epidemical researches, it is usually difficult to hold this assumption. In such fields, usually correlated data occurs. In the presence of correlated data, regarding the dependency between the observations may cause inappropriate analysis results. The objective is to measure the familial relatedness among the family members by using intra and interclass correlation coefficients. These correlation coefficients play an important role in estimating the degree of resemblance among family members with respect to the some characteristics, such as weight, blood pressure, age or cholesterol level. Basically, these correlations are Pearson type correlations, however in genetic studies, the degree of relationship is integrated to the estimators by assigning appropriate weights. A real family data set is used to investigate the familial relatedness among different relative pairs over 20 families with respect to the longitudinal blood pressure measurements. Familial correlations are estimated without any distributional assumptions.

EC701 Room MAL 541 CONTRIBUTIONS TO SURVIVAL ANALYSIS AND RELIABILITY Chair: Takeshi Emura

E0464: **Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures**
Presenter: **Alessandro Beretta**, HEC Liege, Belgium
Co-authors: Cedric Heuchenne
From a survival analysis perspective, bank failures data exhibit heavy censoring rates, but defaults are rare events. This empirical evidence can be explained by the existence of a subpopulation of banks likely immune from bankruptcy. In this regard, we use a mixture cure model to separate the factors with an influence on the susceptibility to default from the ones affecting the survival time of susceptible banks. We extend a semi-parametric proportional hazards cure model to time-varying covariates and we propose a variable selection technique based on its penalized likelihood. By means of a simulation study, we show how this technique performs reasonably well. Finally, we illustrate an application to commercial bank failures in the United States over the period 2006-2016.

E1642: **Estimation and prediction for a distribution with bathtub shape under progressive first failure censoring**
Presenter: **Tanmay Kayal**, Indian Institute of Technologu Patna, India
A two-parameter distribution with bathtub shape under progressive first failure censoring is considered. We obtain maximum likelihood estimates of unknown parameters and then compute observed Fisher information matrix. Further we consider a non linear regression model to estimate

the unknown parameters using least square estimation method and construct 95% Bonferroni confidence intervals for the same. We also estimate the parameters using Bayesian approaches such as Tierney-Kadane method and Metropolis-Hastings algorithm against gamma prior distributions. One- and two-sample prediction problems are discussed under Bayesian framework. We study the performance of proposed methods using Monte Carlo Simulations and finally analyze a real data set for illustration purposes.

E1641: **Bayesian optimum warranty length under Type-II unified hybrid censoring scheme**
Presenter: **Tanmay Sen**, Indian Institute of Technology Patna, India
The determination of optimum warranty length is considered under Type-II unified hybrid censoring scheme. Consumers are willing to purchase a highly reliable product with certain cost constraint. To assure the product reliability and also to remain profitable, the manufacturer provides warranties on product lifetime. Moreover, censored lifetime data are available in practice, to assess the reliability of the product. Therefore, determination of an appropriate warranty length based on censored lifetime data is an important issue to the manufacturer. It is assumed that the lifetime follows a log-normal distribution. We consider a combine free replacement and pro-rata warranty policy (FRW/PRW). The life test is conducted under Type-II unified hybrid censoring scheme. The warranty length is obtained by maximizing an expected utility function.The expectation is taken with respect to the posterior predictive model for time to failure given the available data obtained under Type-II unified hybrid censoring scheme. A real data set is analyzed to illustrate the proposed methodology. We propose a non-linear prorate warranty policy and compare them with linear warranty policy. It is observed that non-linear prorate warranty policy give larger warranty length with maximum profit.

E1786: **A general class of semiparametric accelerated rate models for recurrent event processes under informative censoring**
Presenter: **Sy Han Chiou**, University of Texas at Dallas, United States
Recurrent event analyses in practice often face two challenges: existing model formulations for covariate effects do not fit the data well, and the censoring time is informative after conditioning on covariates. We tackle both challenges in a general class of semiparametric models, which includes proportional rate model, the accelerated rate model, the semiparametric transformation model, and the scale-change model as special cases. Informative censoring is allowed for through a subject-level frailty whose distribution is left unspecified beyond unit expectation. The model parameters are estimated in a two-step procedure. The asymptotic properties of the resulting estimator are established, with the asymptotic variance estimated from a novel resampling strategy. The structure of the proposed model enables model specification tests for each subclass of models, an important issue which has not been well studied. Numerical studies demonstrate that the proposed estimator and tests have attractive performance under both noninformative and informative censoring.

E1628: **A stochastic EM algorithm for maintenance modeling with random improvement factors**
Presenter: **Min Gong**, City University of Hong Kong, Hong Kong
Maintenance policy changes the lifetime behavior of a system by introducing one or more factors into the underlying model, named after improvement factors. Due to many uncontrollable reasons, different maintenance actions should associate with different improvement factors, so these factors should be treated as random. However, random factors lead to great difficulty in parameter estimation. EM algorithm is a standard approach to estimate parameter for model with hidden variables. We compute the expected value of the complete log-likelihood function with respect to the distribution of the hidden variable in the E step, and then maximize it in the M step. The maximum likelihood estimate of parameters can be found by iterating these two steps. However, traditional EM algorithm has the pitfall that the rate of convergence can be painfully slow. Moreover, when the number of simulated samples is large, the maximization step is intractable. To improve the convergence behavior of the Monte Carlo EM algorithm, we here employ the idea of stochastic EM algorithm. Because the expectation in E step is an intractable integral, for certain high-dimensional problems, MCMC simulation is the only known technique capable of providing a solution within a reasonable time. We herein develop some samplers based on the Metropolis-Hastings algorithm, check its feasibility and make comparison.

EG052 Room MAL B30 CONTRIBUTIONS IN REGULARIZATION Chair: Ana Maria Bianco

E1765: **Bayesian non-negative constrained regularised regresssion**
Presenter: **Mauro Bernardi**, University of Padova, Italy
Co-authors: Michele Costola
The aim is to propose a novel Bayesian approach to the problem of variable selection and shrinkage in high dimensional sparse linear regression models when the regression coefficients are also constrained to be positive. The regularisation method is an extension of the Lasso which has been recently cast in a Bayesian framework. Moreover, to deal with the additional problem of variable selection, we propose a Stochastic Search Variable Selection (SSVS) method that relies on a Dirac spike and slab prior where the slab component induces the sparse non-negative regularisation. The non-negative Lasso shrinkage and the SSVS algorithm are then extended to deal with the positive Elastic-Net penalisation and regressor selection. The methodologies are applied to the problem of passive index tracking of large dimensional index in stock markets without short sales. The tracking results indicate that non-negative-Lasso and Elastic Net bases SSVS are successful in asset allocation as compare to the classical solution and alternative Bayesian methods.

E1646: **An iterative sparse-group lasso**
Presenter: **Juan Carlos Laria de la Cruz**, Carlos III University of Madrid, Spain
Co-authors: Rosa Lillo, M Carmen Aguilera-Morillo
Regression models with a sparsity constraint on the solution have become very popular in high dimensional problems. The Sparse-group Lasso (SGL) has gained a lot of attention in last years. Under its simplest formulation, SGL leads to a solution depending on two weight parameters, which control the penalization on the coefficients of the solution. Selecting these weight parameters is still an open problem. In most of the applications of the SGL this problem is left aside, and the parameters are either fixed based on a prior information about the data, or chosen to minimize some error function in a grid of possible values. However, an appropriate election of the parameters deserves more attention, considering that it plays a key role in the structure and interpretation of the solution. In this sense, we present a gradient-free coordinate descent algorithm that automatically selects both penalty parameters of the SGL. The advantages of our approach are illustrated using both real and synthetic data sets.

E1725: **Optimal control and additive perturbations help in estimating ill-posed and uncertain dynamical systems**
Presenter: **Nicolas Brunel**, ENSIIE, France
Co-authors: Quentin Clairon
Ordinary Differential Equations (ODE) are routinely calibrated on real data for estimating unknown parameters or for reverse-engineering. Nevertheless, standard statistical techniques can give disappointing results because of the complex relationship between parameters and states, that makes the corresponding estimation problem ill-posed. Moreover, ODE are mechanistic models that are prone to modeling errors, whose influences on inference are often neglected during statistical analysis. We propose a regularized estimation framework, called Tracking, that consists in adding a perturbation to the original ODE. This perturbation facilitates data fitting and represents also possible model misspecifications, so that parameter estimation is done by solving a trade-off between data fidelity and model fidelity. We show that the underlying optimization problem is an optimal control problem, that can be solved by the Pontryagin Maximum Principle for general nonlinear and partially observed ODE. The same methodology can be used for the joint estimation of finite and time-varying parameters. We show, in the case of a well-specified parametric model, that

our estimator is consistent and reaches the root-n rate. In addition, numerical experiments considering various sources of model misspecifications shows that Tracking still furnishes accurate estimates. Finally, we consider semiparametric estimation on both simulated data and on a real data example.

E1701: **Convergence rates for stochastic inverse problems using variational methods**
Presenter: **Benjamin Sprung**, University of Gottingen, Germany
Co-authors: Thorsten Hohage
Linear inverse problems with Gaussian white noise are considered. In a Hilbert space setting optimal rates of convergence are well known in the literature for Tikhonov regularization and other regularization methods which can be described by a filtered singular value decomposition. Moreover, optimal rates of convergence, also in non-Hilbert norms, have been shown for methods based on wavelet shrinkage. An Advantage of variational methods is that they do not require any knowledge of the forward operator (such as SVD or wavelet-vaguelette decomposition) and they naturally generalize to nonlinear forward operators, Banach space settings and nonquadratic penalty and data-fidelity terms. For a finitely smoothing operator optimal rates of convergence are demonstrated for Besov penalty terms. Extensions to Poisson data with the Kullback-Leibler divergence as data fidelity term are discussed.

E1498: **On regularized regression of categorical responses on categorical predictors**
Presenter: **Ko Sugiura**, Keio University, Japan
Co-authors: Teruo Nakatsuma, Akiyoshi Shimura
Data analysis plays a great important role in healthcare. Data in the field of healthcare is usually collected through a set of questionnaires with categorical scaling. For example, the questionnaire for persons mental health is recorded with the ordinal coding scheme: 0 (no problem), 1 (mild problem), 2 (moderate problem), 3 (severe problem), and 4 (complete problem). Regression analysis for such data has been conducted by regarding the data as continuous on the basis of certain scoring rules for simplicity, because regression on categorical predictors tends to suffer from the high dimensionality of variables. However, such modification makes the meaning of resulting variables abstractive and sometimes ambiguous. For the sake of effective interpretation, one may wish to take the original categorical data into the model without any modification. To deal with the problem of the high dimensionality, we propose a version of the Lasso-type regularization, and demonstrate its effectiveness through an empirical study using the real data on mental stress.

EG024 Room G21A CONTRIBUTIONS TO STATISTICAL MODELLING I Chair: Melanie Birke

E1723: **Semi-parametric estimation in a single index model with endogenous variables**
Presenter: **Melanie Birke**, University of Bayreuth, Germany
Co-authors: Sebastien Van Bellegem, Ingrid Van Keilegom
A semiparametric single-index model is considered where endogeneity is present in the explanatory variables. The presence of an instrument is assumed, that is noncorrelated with the error term. Endogeneity is a central issue when modeling statistical data coming from human or medical sciences and occurs when some of the independent variables in a regression model are correlated with the error term. It can arise when relevant explanatory variables are omitted from the model, as a result of sample selection errors or when unobserved subject selection occurs in experimental studies. When endogeneity is present, ordinary regression techniques produce biased and inconsistent estimators. A possible way out is to make use of the so-called instrumental variables. This approach is combined with the single-index model which is often used for dimension reduction. An estimator of the parametric component of the model is proposed, which is the solution of an ill-posed inverse problem. In a first step the unknown link function is estimated with kernel methods. In a second step this estimator is used to minimize a criterion function over a class of parameters. The estimator is shown to be asymptotically normal under certain regularity conditions common in empirical process theory. A simulation study is conducted to illustrate the finite sample performance of the proposed estimator.

E0569: **A point process model for pulses of the DPT from an image**
Presenter: **Inger Fabris-Rotelli**, University of Pretoria, South Africa
Co-authors: Alfred Stein
The Discrete Pulse Transform (DPT) is a multiscale decomposition of an image obtained using the LULU smoothers. The LULU smoothers are nonlinear and have many powerful properties as such edge and total variation preservation. They have been developed for multiple dimensions here we focus on images. A point pattern is extracted from an image using the DPT and modelled as a spatially clustered pattern with trend in both the rows and columns of the image. The clustered nature arises naturally from the images structures which are well extracted by the DPT, and is confirmed prior to modelling. The fitted models perform well and extracts the spatial intensity of the image. The DPT provides a representation of an image at all scales from detail up to background shading - and enables informed analysis of the content and features of an image. The aim is to expand the DPT image application into the spatial domain, a natural domain for image structures having obvious spatial dependence.

E1337: **Box-Cox transformation for regression models with random effects**
Presenter: **Amani Almohaimeed**, Durham University, United Kingdom
Co-authors: Jochen Einbeck
Regression analyses can help us to detect trends, examine relationships and draw meaningful conclusions from experimental data. However, the assumptions of normality and of homoscedasticity of the response must be fulfilled prior to starting to analyze the data. The aim is to ensure the validity of a normal response distribution using the Box-Cox power transformation. The extension of this transformation to the linear mixed effects model was previously proposed in the case of a Gaussian random effect distribution. An obvious concern of assuming a normal random effect distribution is whether there are any harmful effects of a potential misspecification. This problem can be avoided by estimating this distribution implicitly via the EM algorithm, through the use of a technique known as Nonparametric Maximum Likelihood, which, for our purposes, is adapted towards a Nonparametric Profile Maximum Likelihood technique. The feasibility of the approach is demonstrated through examples using a new R package, boxcoxmix, to be available on the Comprehensive R Archive Network (CRAN).

E1558: **Modeling the ARMA random effects covariance matrix in logistic random effects models**
Presenter: **Keunbaik Lee**, Sungkyunkwan University, Korea, South
Logistic random effects models (LREMs) have been frequently used to analyze longitudinal binary data. When using a random effects covariance matrix to make proper inferences on covariate effects, the random effects in the models account for both within-subject association and between-subject variation. Estimation of the covariance matrix is challenging, however, because it is high-dimensional and should be positive definite. To overcome these limitations, two Cholesky decomposition approaches were proposed for precision matrix and covariance matrix: modified Cholesky decomposition and moving average Cholesky decomposition, respectively. When there are non-trivial and complicated correlations of repeated outcomes, however, the two approaches may not work. We combine the two decomposition approaches to model the random effects covariance matrix in the LREMs, thereby capturing a wider class of sophisticated dependence structure while achieving parsimony in parametrization. We then analyzed lung cancer data using our proposed model.

E1733: **Smoothing the logistic model**
Presenter: **Hugo Maruri**, QMUL, United Kingdom
Co-authors: Maria Vazquez
Smooth supersaturated polynomials have been used for building emulators in computer experiments. The response surfaces built with this method are simple to interpret and have spline-like properties. We extend the methodology to build smooth logistic regression models. The approach we follow is to regularize the likelihood with a penalization term that accounts for the roughness of the regression model. The response surface follows data closely yet it is smooth and does not oscillate. We illustrate the method with simulated data. We also present a recent application to build a prediction rule for psychiatric hospital readmissions of patients with a diagnosis of psychosis. This application uses data from the OCTET clinical trial.

EG018 Room MAL 151 CONTRIBUTIONS IN NONPARAMETRIC METHODS Chair: Mayer Alvo

E1589: **Combined permutation tests for linear regression models**
Presenter: **Stefano Bonnini**, University of Ferrara, Italy
In the econometric literature, several contributions have been dedicated to permutation solutions for linear model selection. A permutation version of the ANOVA test for a linear regression model with q explanatory variables might be based on the F test statistic. Its null permutation distribution might be obtained by permuting the vector of observed values of the dependent variable and keeping the columns of regressors fixed. As alternative, we propose a solution based on the combination of the partial permutation tests on the single regression coefficients. This method has the advantage that the overall test on the full model is broken down into q sub-problems and may be intended as a multiple test. By means of a suitable correction of the partial p-values, it allows the detection of the relevant predictive variables avoiding the inflation of the type I error rate of the overall test. Thus it is a nonparametric alternative to the backward elimination method or other stepwise procedures, with two main advantages: it is time saving and it controls the type I error probability. Good power behaviour and utility of the method are shown through simulations and application to a real problem.

E1745: **Use of the two-dimensional Kolmogorov-Smirnov test to measure spatial concentration in geospatial data**
Presenter: **Takaaki Ohnishi**, The University of Tokyo, Japan
Co-authors: Takayuki Mizuno, Tsutomu Watanabe
Since micro-geographic data is becoming available, it is necessary to develop robust methods applicable to such data. Industry concentration is an universal property observed in most countries and at various spatial scales. The spatial concentration of industries has been traditionally measured by using cluster-based methods which is defined on a discrete definition of space. In these methods, space is divided into subunits, so the position of the boundaries and level of observation have an impact resulting in the modifiable areal unit problem. In order to avoid this problem, recent studies have applied distance-based methods which consider space as continuous. Although these methods allow an exact and unbiased analysis of the spatial concentration at all scales simultaneously, it is necessary to define the values of some parameters and the shape of the kernel function. To overcome these disadvantages, we present the use of the two-dimensional Kolmogorov-Smirnov two sample test to measure spatial concentration. The Kolmogorov-Smirnov test computes the maximum difference between two cumulative distribution functions yields a P-value. The spatial concentration can be characterized by P-value. We illustrate the proposal by analyzing Japanese telephone directory data which contains about 7 million establishments with latitude, longitude, and industrial sector information.

E1494: **Tree-based regression for interval-valued data**
Presenter: **Yan Sun**, Utah State University, United States
Co-authors: Chih-Ching Yeh, Adele Cutler
Regression methods for interval-valued data have been increasingly studied in recent years. As most of the existing works focus on linear models, it is important to note that many problems in practice are nonlinear in nature and therefore development of nonlinear regression tools for interval-valued data is crucial. We propose a tree-based regression method for interval-valued data, which is well applicable to both linear and nonlinear problems. Unlike linear regression models that usually require additional constraints to ensure positivity of the predicted interval length, the proposed method estimates the regression function in a nonparametric way, so the predicted length is naturally positive without any constraints. A simulation study is conducted that compares our method to popular existing regression models for interval-valued data under both linear and nonlinear settings. Furthermore, a real data example is presented where we apply our method to analyze price range data of the Dow Jones Industrial Average index and its component stocks.

E0400: **New nonparametric estimation of entropy and applications**
Presenter: **Issam El Hattab**, ENCG-Casablanca, Hassan 2 University, Morocco
The focus is on the nonparametric estimation of the Shannon's entropy and the some statistical tests. We introduce two new families of entropy estimators based on smooth quantile density estimators. The consistency and asymptotic normality of the proposed estimates are obtained. Our arguments, used to establish the asymptotic normality of our estimator, make use of an original application of the strong invariance principle for the uniform empirical process. A simulation study for the comparison, in terms of mean squared error, is performed, from which it follows that the proposed entropy estimates provides in this sense the most efficient approximation. Finally, we give some novel statistical tests based on our estimates.

E1605: **How many point masses do we need for non-parametric deconvolution and maximum likelihood mixture densities?**
Presenter: **Timothy Hyndman**, The University of Melbourne, Australia
Co-authors: Peter Taylor, Aurore Delaigle
In non-parametric density deconvolution problems where the error distribution is unknown, one must solve an optimization problem to find a discrete probability distribution that approximates a continuous target distribution. In practice, this discrete distribution has surprisingly few point masses. The same phenomenon can also be observed when finding maximum likelihood mixture densities. The goal is to try to understand why this is happening, and to be able to say something about the number of point masses that will be required prior to performing the optimization. Some basic results for the maximum likelihood mixture problem will be presented by viewing the problem as an optimization problem, rather than a statistical one.

EG654 Room CLO 204 CONTRIBUTIONS IN INCOMPLETE DATA Chair: Clemence Leyrat

E1601: **An EM-type algorithm for maximum likelihood estimation of spatial models with positional errors**
Presenter: **Marco Bee**, University of Trento, Italy
Co-authors: Giuseppe Espa, Diego Giuliani, Maria Michela Dickson, Emanuele Taufer, Flavio Santi
Positional errors are often encountered in spatial statistics. The coordinates c_i^* of not properly located observations can be written as $c_i^* = c_i + m_i$, where c_i are the true (but unobserved) coordinates and m_i is a bivariate normal random vector which models the error. Given that the m_is can be interpreted as missing data, an estimation method based on the EM algorithm is developed. The impact of various hypotheses on the parameters of m (zero or non-zero mean, heteroskedasticity and/or dependence of the two components) is studied via simulation. The details of the algorithm are worked out in the framework of the Spatial Lag model, with a connectivity matrix defined as a non-linear function of the Euclidean distance between the observations, but can easily be extended to other spatial statistics models. From the technical point of view, both the E and the M step are not in closed form. A possible solution consists of computing the E-step by Monte Carlo simulation and using standard numerical optimization routines for the M-step.

E1624: **Robust assessment of cross-over designs against missing values**
Presenter: **Ed Godolphin**, University of Surrey, United Kingdom
Co-authors: Peter Godolphin
In scientific experiments where human behaviour or animal response is intrinsically involved, such as clinical trials, there is a strong possibility that some observations will not be recorded. Missing data in a clinical trial has the potential to impact severely on study quality and precision of estimates. In studies which use a cross-over design, even a small number of missing values can lead to the eventual design being disconnected. In this case, some or all of the treatment contrasts under test cannot be estimated and the experiment is compromised since little can be achieved from it. We consider experiments that use a cross-over design. Methods to limit the impact of missing data on study results are explored. It is shown that maximal robustness and perpetual connectivity of the planned design are properties which are related and which guard against the possibility of a disconnected eventual design. A procedure is proposed which assesses planned designs for robustness against missing values and the method is illustrated by assessing several designs that have been considered in the recent literature on cross-over studies.

E1759: **Distance estimation for mixed continuous and categorical data with missing values**
Presenter: **Eduardo Mendes**, Fundacao Getulio Vargas, Brazil
Co-authors: Glauco Azevedo
A methodology is proposed to estimate the pairwise distance between mixed continuous and categorical data with missing values. Distance estimation is the base for many regression/classification methods, such as nearest neighbors and discriminant analysis, and for clustering techniques such as k-means and k-medoids. Classical methods for handling missing data rely on mean imputation, that could underestimate the variance, or regression-based imputation methods. Unfortunately, when the goal is to estimate the distance between observations, data imputation may perform badly and bias the results toward the data imputation model. We estimate the pairwise distances directly, treating the missing data as random. The joint distribution of the data is approximated using a multivariate mixture model for mixed continuous and categorical data. We present an EM-type algorithm for estimating the mixture and a general methodology for estimating the distance between observations. Simulations shows an improved performance of our method when compared to traditional imputation methods.

E0916: **ROC surface analysis in presence of verification bias**
Presenter: **Khanh To Duc**, University of Padova, Italy
Receiver operating characteristic (ROC) surface and volume under ROC surface (VUS) are frequently used for assessing accuracy of a diagnostic test when disease status has three categories (e.g., non-diseased, intermediate and diseased). A problem encountered in studies that evaluate a new diagnostic test is that often not all patients undergo disease verification, due to the expensiveness and invasiveness of the gold standard test. Using only the verified patients to estimate the ROC surface and its volume would result in verification bias. We present several verification bias-corrected estimators of the ROC surface and the VUS, based on imputation and re-weighting techniques. These estimators work when the true disease status is missing at random . In addition, we present an R package and the corresponding Shiny web application, which support researchers in carrying out the ROC surface analysis in presence of verification bias.

E0498: **Computationally efficient sequential regression imputation for multilevel datasets**
Presenter: **Murat Recai Yucel**, State University of New York, United States
Co-authors: Tugba Akkaya-Hocagil
An alternative approach based on variable-by-variable imputation is considered when either the joint model is nonsensical due to non-applicability or due to high dimensionality. Joint models can be nonsensical in situations when survey item can only be applicable to certain sub-groups. We improve computational aspects of the previous methods relying on traditional sampling techniques such as Gibbs sampler under linear mixed-effects models. In particular, we eliminate the adverse effect of increased inter-dependency on random-effects and its impact on convergence. The improved algorithm greatly reduces computation times in the context of variable-by-variable imputation in surveys or other data problems with large number of variables subject to missingness. We present a comprehensive simulation study as well as real data example from nationally-conducted survey.

EG249 Room MAL 402 CONTRIBUTIONS IN STATISTICAL METHODS FOR ECONOMICS AND FINANCE Chair: Jonas Andersson

E1741: **Correlations between irregularly spaced time series**
Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway
The problem of analyzing relations between time series observed irregularly and at different points in time is addressed. In particular, computing covariances and correlations and the extension to linear regression is treated. The first difference of processes which are integrated of order one, I(1)-processes, are of special interest because they are often used as models for stock returns. The covariances of stock returns, in turn, is interesting since they are used in portfolio optimization. Different interpolation methods and modeling the irregularly spaced observations explicitly are investigated as well as simulation based methods. For the case where the return process and the process governing the times of trade are dependent, an unbiased estimation method is proposed and investigated.

E1645: **Backtesting in finance and prequential analysis**
Presenter: **Hideatsu Tsukahara**, Seijo University, Japan
Prequential analysis provides us with a suitable framework for validating backtesting procedures. In prequential analysis, data are observed sequentially in time, and at each time it is required to make a probability forecast related to the next observation based on the past and current data (thereby the name "prequential"). Given a sequence of probability forecasts, there always exists a unique probability measure on the space of all sequences; it is called a probability forecasting system by Dawid. It follows from Blackwell-Dubins theorem that if two probability forecasting systems are equivalent, then the total variation distance between the corresponding probability forecasts converges to 0 a.s.; in fact, the total

variation distance between the conditional distributions for the whole future given the past converges to 0, a.s. From the Bayesian point of view, it may be interpreted to imply that the opinions of two individuals will merge after a sufficiently large number of observations under the assumption of equivalence of their subjective probabilities. We will study condition for the absolute continuity using the martingale theory. Their implications for the validation of backtesting procedures are examined with reference to Mark Davis' calibration concept. A few simple examples will be shown to exemplify the theory. Finally, possible extensions will be discussed.

E1267: **Sentiment spillover effects for US and European companies**
Presenter: **Anastasija Tetereva**, University St Gallen, Switzerland
Co-authors: Francesco Audrino
The fast-growing literature on the news and social media analysis provide empirical evidence that the financial markets are often driven by information rather than facts. However, the direct effects of sentiments on the returns are of main interest. We propose to study the cross-industry influence of the news for a set of US and European stocks. The graphical Granger causality of the news sentiments - excess return networks is estimated by applying the adaptive Lasso procedure. We introduce two characteristics to measure the influence of the news coming from each sector and analyze their dynamics for a period of 10 years ranging from 2005 to 2014. The results obtained provide insight into the news spillover effects among the industries and the importance of sentiments related to certain sectors during periods of financial instability.

E1686: **Regular paths in financial markets: investigating the Benford's Law**
Presenter: **Jessica Riccioni**, University of Macerata, Italy
Co-authors: Roy Cerqueti
The aim is to verify whether the Benford's law is valid in the context of global stock market and for all the indexes composing it, with the intention of assessing the presence of data regularities and identifying discrepancies. Specifically, we check the reliability of Benford's Law for all indexes listed on the stock exchanges of the various countries whose time series are available, with a particular reference to prices and volumes of stocks. To pursue our scopes, we adopt comparison criteria grounded on statistical theory, like the Chi-squared test and the mean of the absolute deviations for both the distributions of the first and the second meaningful digits. Evidence of violations are provided, and some insights taken from the historical facts and economic shocks are carried out.

E1390: **Bias calibration for robust estimation of inequality indices in small areas**
Presenter: **Setareh Ranjbar**, University of Geneva, Switzerland
Co-authors: Elvezio Ronchetti
Today the availability of rich sample surveys provides a ground for researchers and policy makers to pursue more ambitious objectives. This information in line with auxiliary data coming through administrative channels is used for a better prediction/estimation of social and economic indices, e.g. inequality or poverty measures, that can help to determine more precisely their target domains. The domains for which the sample size is not large enough to provide an acceptable direct estimate, are referred to as small areas. The existence of outliers in the sample data can significantly harm the estimation for areas in which they occur, especially where the domain-sample size is small. Robust estimation of finite population total and mean in the presence of outliers has been discussed in the literature and the results has been extended to the cumulative distribution time ago. Based on a robust EBLUP estimation, we propose two new approaches to calibrate for the bias of nonlinear functionals, such as the Gini index and when the so-called representative outliers come from a skewed heavy tail distribution. The method is also used to impute missing income values, a common occurrence in labour force surveys.

EG030 Room MAL 415 CONTRIBUTIONS IN HIGH DIMENSIONAL STATISTICS Chair: Cristian Gatu

E1411: **Adaptively weighted group Lasso for semiparametric quantile regression models**
Presenter: **Toshio Honda**, Hitotsubashi University, Japan
Co-authors: Ching-Kang Ing, Wei-Ying Wu
An adaptively weighted group Lasso procedure is proposed for simultaneous variable selection and structure identification for varying coefficient quantile regression models and additive quantile regression models with ultra-high dimensional covariates. Under a strong sparsity condition, we establish selection consistency of the proposed Lasso procedure when the weights therein satisfy a set of general conditions. This consistency result, however, is reliant on a suitable choice of the tuning parameter for the Lasso penalty, which can be hard to make in practice. To alleviate this difficulty, we suggest a BIC-type criterion, which we call high-dimensional information criterion (HDIC), and show that the proposed Lasso procedure with the tuning parameter determined by HDIC still achieves selection consistency. Our simulation studies support strongly our theoretical findings.

E1658: **Polynomial shrinkage of large-dimensional covariance matrices**
Presenter: **Nicolas Tavernier**, KU Leuven, Belgium
Co-authors: Geert Dhaene
An optimal rule is derived for shrinking large-dimensional sample covariance matrices under Frobenius loss. The rule generalizes Ledoit and Wolf's optimal linear shrinkage rule to broader parametric families of rules. The families include, for example, polynomial, piecewise linear, and spline rules. The oracle version of the optimal rule is very simple and attains the lower bound on the Frobenius loss in finite samples. Feasible versions approximately attain the lower bound under large-dimensional asymptotics where $p/n \to c > 0$, but are not generally available in closed form. The polynomial family of rules, however, admits a closed-form estimator of the optimal rule. It is derived using results from random matrix theory and an algorithm to calculate Wishart moments of arbitrary order. In settings that have been studied earlier, polynomial shrinkage is found to substantially reduce the Frobenius loss compared to linear shrinkage. Polynomial shrinkage is conceptually easy, does not require non-convex optimization in high dimension, and allows $p > n$.

E1612: **Honest data-adaptive inference for the average treatment effect using penalised bias-reduced double-robust estimation**
Presenter: **Vahe Avagyan**, Ghent University, Belgium
Co-authors: Stijn Vansteelandt
The presence of confounding by high-dimensional variables complicates estimation of the average effect of a point treatment. On the one hand, it necessitates the use of variable selection strategies or more general data-adaptive high-dimensional statistical methods. On the other hand, the use of such techniques tends to result in biased estimators with a non-standard asymptotic behaviour. Double-robust estimators are vital for offering a resolution because they possess a so-called small bias property. This means that their bias vanishes faster than the bias in the nuisance parameter estimators when the relevant smoothing parameter goes to zero, making their performance less sensitive to smoothing. This property has been exploited to achieve valid (uniform) inference of the average causal effect when data-adaptive estimators of the propensity score and conditional outcome mean both converge to their respective truths at sufficiently fast rate. We extend this work in order to retain valid (uniform) inference when one of these estimators does not converge to the truth, regardless of which. This is done by generalising prior work for low-dimensional settings to incorporate regularisation. The proposed penalised bias-reduced double-robust estimation strategy exhibits promising performance in extensive simulation studies and a data analysis, relative to competing proposals.

E1694: **Robust sieve M-estimation with an application to dimension reduction**
Presenter: **Julien Bodelet**, University of Geneva, Switzerland
Co-authors: Davide La Vecchia
Non/semiparametric techniques are routinely applied in the analysis of high dimensional data. Robust procedures for such techniques are highly requested, since the huge amount of records can make outlier detection very difficult. We explain how to achieve robust dimension reduction using sieve M-estimators, which represent a general and flexible class of estimators for non/semiparametric models. First, by means of the von Mises calculus for statistical functionals and the empirical process theory, a new theoretical device, called sieve Influence Function, suitable to characterize the class of infinitesimally robust sieve M-estimators, is introduced. Then, rates of convergence of the robust estimators on non-convex sieve spaces are derived: this extends the asymptotic results available for sieve M-estimators on convex sieve spaces. Finally, the use of the estimation procedure is illustrated for the class of Dynamic Semiparametric Factor Models, which are applied for dimension reduction in the analysis of functional Magnetic Resonance Imaging (fMRI) data. Monte Carlo simulations illustrate that the robust M-estimators outperform classical sieve M-estimators in the presence of outliers. An application to real fMRI data is also shown.

E1834: **Sequential change-point detection in panel data models**
Presenter: **Jaewon Huh**, University of Texas Southwestern Medical Center, United States
Co-authors: Sangyeol Lee
A sequential monitoring method is developed for linear panel data models based on the maximum eigenvalue of the empirical covariance matrix. Panel Data models are known to play a major role in analysis of high-dimensional data owing to its similarity to time series models. We pay special attention in extending the change-point detection techniques to the high-dimensional settings with theoretical backup. The performance of the proposed method is evaluated through a simulation study with various parameter settings. Our findings show that the proposed method performs adequately. Moreover, the use of this method is theoretically well equipped.

CI288 Room Beveridge Hall NON-LINEAR AND NON-STATIONARY MODELS Chair: Bent Nielsen

C0167: **Functional-coefficient cointegrating regression with endogeneity**
Presenter: **Qiying Wang**, University of Sydney, Australia
Nonparametric estimation of functional-coefficient cointegrating regression models is explored where the structural equation errors are serially dependent and the regressor is endogenous. Generalizing earlier works, the self-normalized local kernel and local linear estimators are shown to be asymptotic normal and to be pivotal upon an estimation of co-variances. Our new results open up inference by conventional nonparametric methods to a wide class of potentially nonlinear cointegrated relations.

C0168: **Threshold stochastic unit root models**
Presenter: **Jesus Gonzalo Munoz**, Universidad Carlos III de Madrid, Spain
The main goal is to propose a model that is able to explain the existence of unit roots in economic variables. This is done by introducing a Threshold Stochastic Unit Root Model (TARSUR), where the largest root of an AR is allowed to vary between regimes triggered by a threshold variable, in such a way that the expectation of that root is still unity. The stationarity properties of the model are analyzed, and a testing strategy is presented. Applications to stock prices, commodity prices, interest rates and exchange rates are shown.

C0169: **Regressions with fractional d=0.5 and weakly nonstationary processes**
Presenter: **Ioannis Kasparis**, University of Cyprus, Cyprus
Co-authors: James Duffy
Despite major advances in the statistical analysis of fractionally integrated time series models, no limit theory is available for sample averages of fractionally integrated processes with memory parameter d=0.5. We provide limit theory for sample averages of a general class of nonlinear transformations of fractional d=0.5 (I(0.5)) and Mildly Integrated (MI) processes. Although I(0.5) processes lie in the nonstationary region, the asymptotic machinery that is routinely used for I(d), d>1/2 is not valid in the I(0.5) case. In particular, FCLTs are not applicable to I(0.5) processes and a different approach is required. A general method that applies to both I(0.5) and MI processes is proposed. We show that sample averages of transformations of I(0.5) and MI processes converge in distribution to the convolution of the transformation function and some Gaussian density evaluated at a possibly random point. The type of nonlinear transformations under consideration accommodates a wide range of regression models used in empirical work including step type discontinuous functions, functions with integrable poles as well as integrable kernel functions that involve bandwidth terms. Our basic limit theory is utilised for the asymptotic analysis of the LS and the Nadaraya-Watson kernel regression estimator.

C0430: **Testing for autocorrelation in non-linear and non-stationary regressions**
Presenter: **Vanessa Berenguer Rico**, University of Oxford, United Kingdom
Co-authors: Yingyu Guo, Bent Nielsen
An auxiliary-regression-test is analyzed for autocorrelation of residuals in non-linear and non-stationary regressions. It is shown that the test statistic is asymptotically chi-square distributed under the null of no autocorrelation. Monte Carlo simulations reveal that the test statistic has good finite sample properties.

CO500 Room Bloomsbury BUSINESS CYCLES AND MACROECONOMIC IMBALANCES Chair: Lola Gadea

C0507: **Regional business cycles in Europe: Dating and clustering**
Presenter: **Eduardo Bandres**, Universidad de Zaragoza, Spain
Co-authors: Ana Gomez-Loscos, Lola Gadea
The aim is threefold. First, we examine the business cycles of European regions and the comovements among them. Second, using Finite Mixture Markov models, we compute regional business cycle datings that allow us to identify clusters of European regions that show similar business cycle features. Third, we develop a new index to measure within-country homogeneity. We find that comovements among regions is, on average, quite low. However, we find a progressive increase of the spatial correlation during the convergence process and a sharp rise during the Great Recession. We identify five different groups of European regions: a Greek group, a southern Germany group, a core group, a northern group and the largest group, which encompasses the rest of the regions. We also find sizable border effects.

C0540: **External imbalances and recoveries**
Presenter: **Cecilio Tamarit**, University of Valencia, Spain
Co-authors: Lola Gadea, Mariam Camarero, Ana Gomez-Loscos
The purpose is analyzing the role that macroeconomic external imbalances, such as the net investment position or the current account balance, may have on the depth, duration, speed and shape of recoveries. Differences in the speed of recoveries are, in principle, related to the interplay between macroeconomic and structural factors. In a first step, we dissect recoveries for a wide sample of OECD countries and extract their main

cyclical features, which may have dramatic consequences in terms of long-term growth, job creation and economic wealth. Secondly, we link these measures with several indexes of external imbalances Additionally, we study whether the degree of openness, its interaction with the current account balance or real exchange rates are able to explain the asymmetries in the recoveries for different groups of countries. Our findings aim at identifying different groups of countries according to characteristics of the and how they are affected by different levels of external imbalances.

C0756: **Finite Markov mixture modeling to cluster turning points**
Presenter: **Ana Gomez-Loscos**, Bank of Spain, Spain
Co-authors: Maximo Camacho, Lola Gadea
A new date-then-average approach is proposed to date business cycle turning points from a large set of economic indicators. Each individual turning point date is viewed a realization of a mixture of an unspecified number of separate Gaussian distributions, which determine the number and the dates of the aggregated turning points of the reference cycle. Finite mixture models techniques are used to determine the number of turning points of the reference cycle, to estimate the parameters for the separate distributions, to determine the mixing proportions and to perform statistical inference for the estimated reference cycle. By means of a Monte Carlo experiment, we show the high accuracy of the method to determine turning points. Using a set of US economic indicators, we show that the method is able to identify the NBER turning points with a maximum deviation of one quarter.

C0664: **Exchange rate regime and external adjustment: An empirical investigation for the U.S.**
Presenter: **Alberto Fuertes**, Bank of Spain, Spain
The relationship between the U.S. net external position and the exchange rate regime is analyzed. We find a structural break in the U.S. net external position at the end of the Bretton Woods system of fixed exchange rates that changed both the mean and variance of the series. On average, the U.S. changed from a creditor to a debtor position and the variance of the external position increased during the floating period. This increase is to a large extent due to the valuation component of external adjustment, that accounts for 54% of the variance of the U.S. external position during the floating period but only 29% during the fixed exchange rate period. Further analysis shows that the exchange rate regime mainly affects the valuation channel of external adjustment. There is also evidence of another structural break in the U.S. external position around the time of the introduction of the euro. Finally, we document asset pricing implications from the relationship between the exchange rate regime and the external adjustment process, as external imbalances predict the foreign exchange once the exchange rate regime is taken into account.

C0835: **Growth and business cycle in Argentina: A long-run approach**
Presenter: **Isabel Sanz**, University of Valencia, Spain
Co-authors: Lola Gadea
The study of the long-term evolution of the Argentine economy and the determination of the moment in which it began its estrangement with other developed economies, such as Australia and Canada, constitutes the central axis of the historiography of this country. However, the added problem presented by the Argentine economy is its volatility. For this reason, the long term has to be influenced by the short term, which requires a more detailed study of the cyclical behavior. The objective is precisely to descend into the short term to explain its relative long-term evolution given the characteristics and peculiarities of the economic cycle. For this, beyond the mere analysis of synchronization, new measures are proposed additional to those used in the traditional analyzes of cycles to study the duration, the depth, the amplitude and the form of the same ones. The results obtained point to a cyclical synchronization that goes in parallel to the long-term evolution and, therefore, explains Argentina's convergence process with Australia and Canada.

CO244 Room Chancellor's Hall TESTING AND FORECASTING FINANCIAL TIME SERIES Chair: Jeroen Rombouts

C1154: **Loosening the chains on multi factor asset pricing models: Letting time varying betas speak**
Presenter: **Fabio Calonaci**, Queen Mary University of London, United Kingdom
Co-authors: George Kapetanios, Richard Baillie
Estimates of the Fama MacBeth version of the 3 Factor model are derived by implementing some recent, locally smoothed, time varying parameter methods which are non parametric. Importantly, the approach selects the bandwidth on the basis of cross validation of estimating the 3 Factor model. Another novelty of the approach allows computation of time variation in the factor loadings, and avoids heavy theoretical restrictions on the model. Once we make these allowances on the specification of the 3 Factor model, we see there is a new level of explanatory power in the model both in terms of time series and cross sectional variation in the expected returns. There are also further insights into the asset pricing anomalies such as size, value and momentum.

C1211: **Limit order book liquidity and moments of the return distribution**
Presenter: **Tolga Cenesizoglu**, HEC Montreal and LSE, Canada
The relationship between liquidity provision through limit orders and the moments of the return distribution is analyzed. Assuming exogenous random order arrivals, we predict that (i) volatility increases in ask and bid side illiquidity (as measured by the price impact parameter) (ii) skewness increases in illiquidity imbalance and (iii) kurtosis increases in the convexity of the price impact function. We verify these relations using 11 year of comprehensive NYSE limit order book (LOB) data. We establish causality by examining the impact of regulatory shocks to LOB liquidity provision on the moments of returns using diff-in-diff analyses.

C1140: **Individual stock variance premia properties**
Presenter: **Jeroen Rombouts**, ESSEC Business School, France
Co-authors: Francesco Violante, Lars Stentoft
For the individuals stock of the S&P 500 index, the aim is to estimate the Variance Risk Premium, defined as the difference between risk neutral and physical expectations of an asset's total return variation which has market return predictability and is of fundamental importance for validation and development of new asset pricing models. Variance swap payoffs are highly volatile time series, with time varying variance levels and extreme payoffs during volatile market conditions, and to extract the VRP we use signal extraction techniques based on a state-space representation of the model and the Kalman-Hamilton filter. Our proposed approach provides measurement error free estimates of the part of the VRP related to normal market conditions, and allows constructing variables indicating agents' expectations under extreme market conditions. The framework only requires option implied volatility, e.g. the VIX index for the S&P 500, data and daily returns for the underlying, the sources of which are free and readily available for many assets.

C1213: **Pricing individual stock options using both stock and market index information**
Presenter: **Lars Stentoft**, University of Western Ontario, Canada
Co-authors: Jeroen Rombouts, Francesco Violante
When it comes to individual stock option pricing, most, if not all, applications consider a univariate framework in which the dynamics of the underlying asset is considered without taking the evolution of the market or any other risk factors into consideration. From a theoretical point of view this is clearly unsatisfactory as we know, i.e. from the Capital Asset Pricing Model, that the expected return of any asset is closely related to

the exposure to the market risk factors. On top of this theoretical inconsistency in empirical applications it is often difficult to precisely asses and appropriately measure risk premia from individual stock returns alone. We model the evolution of the individual stock returns together with market index in a bivariate model that allows us to estimate risk premia in line with the theory. We assess the performance of the model to price individual stock options on the constituent stocks in the Dow Jones Industrial Average.

C1332: **Testing for jump spillovers without testing for jumps**
Presenter: **Marcelo Fernandes**, Sao Paulo School of Economics, FGV, Brazil
Co-authors: Walter Distaso, Valentina Corradi
Statistical tools are developed for testing conditional independence among the jump components of the daily quadratic variation, which we estimate using intraday data. To avoid sequential bias distortion, we do not pretest for the presence of jumps. If the null is true, our test statistic based on daily integrated jumps weakly converges to a Gaussian random variable if both assets have jumps. If instead at least one asset has no jumps, then the statistic approaches zero in probability. We show how to compute asymptotically valid bootstrap-based critical values that result in a consistent test with asymptotic size equal to or smaller than the nominal size. Empirically, we study jump linkages between US futures and equity index markets. We find not only strong evidence of jump cross-excitation between the SPDR exchange-traded fund and E-mini futures on the S&P 500 index, but also that integrated jumps in the E-mini futures during the overnight period carry relevant information.

CO522 Room G11 COMMODITY MARKETS: PRICING AND TRADING Chair: Ana-Maria Fuertes

C0518: **On risk-neutral skewness and commodity pricing**
Presenter: **Weiqing Tang**, University of Birmingham, United Kingdom
Co-authors: Ana-Maria Fuertes, Zhenya Liu
The predictive content of risk-neutral skewness for the dynamics of commodity futures prices is investigated. A risk-neutral implied skewness measure obtained from weekly 10-year options and futures return data is shown to explain the variability of time series futures return, and to price the cross-section of commodity futures return. The higher (lower) the current model-free implied skewness, the higher (lower) the subsequent return. Under pricing (Over pricing) implied by positive (negative) skewness document mispricing correction procedure behind option implied skewness measure. Trading portfolio strategy sorted by risk-neutral skewness outperform its counterpart sorted by realized skewness with extra 14.5% annual return. Results are robust to several alternatives.

C1320: **Harvesting commodity styles: An integrated framework**
Presenter: **Ana-Maria Fuertes**, Cass Business School - City University London, United Kingdom
Co-authors: Adrian Fernandez-Perez, Joelle Miffre
A portfolio allocation framework is developed to study the benefits of style integration and to compare the effectiveness of alternative integration methods in commodity markets. The framework is flexible enough to be applicable to any asset class for either long-short, long- or short-only styles. We study the nave equal-weighted integration and sophisticated integrations where the style exposures are sample-based by utility maximization, style rotation, volatility-timing, cross-sectional pricing or principal components analysis. Considering the universe of eleven long-short commodity styles, we document that the nave integration enhances each of the individual styles in terms of their reward-to-risk tradeoff and crash risk profile. Sophisticated integrations do not challenge the nave integration and the rationale is that, while also achieving multiple-style exposures, the equal-weighting approach circumvents estimation risk and perfect-foresight bias. The findings hold after trading costs, various reformulations of the sophisticated integrations, economic sub-period analyses and data snooping tests inter alia.

C1445: **On what drives commodity returns: Market, sector or idiosyncratic factors**
Presenter: **Andrew Vivian**, Loughborough University, United Kingdom
Co-authors: Jun Ma, Mark Wohar
The relationship between 43 commodity returns is examined by using a dynamic factor model with time varying stochastic volatility. The dynamic factor model decomposes each commodity return into a market, sector-specific and commodity-specific component. It enables the variance attributed to each component to be estimated at each point in time. We find the return variation explained by the common factor has increased substantially for the recent period and is statistically significant for the vast majority of commodities since 2004 (at each point in time). This phenomenon is strongest for non-perishable products. We link the amount of variation explained by the common factor to economic variables.

C1424: **The formation of FFA rates in dry bulk shipping: Spot rates, risk premia, and heterogeneous expectations**
Presenter: **Ioannis Moutzouris**, Cass Business School, United Kingdom
Co-authors: Nikos Nomikos
The formation of FFA rates in the dry bulk shipping industry is examined. We illustrate that the bulk of volatility in the FFA basis can be attributed to expectations about future physical market conditions rather than expectations about future risk premia, as is commonly suggested in the commodity markets literature. Despite this finding, though, there appears to be a bias in FFA rates in the form of both a strong momentum effect and significant predictability of risk premia by lagged price-based signals and economic variables that reflect recent changes in the physical market conditions. We further contribute to the literature by developing an asset pricing framework that can explain both the existence of momentum and the documented sort of predictability of future risk premia. The distinct feature of our framework is that, apart from having as is standard in the literature different objective functions, agents might also differ in the way they form expectations about future market conditions. Accordingly, we illustrate formally that, to simultaneously match the observed empirical regularities, one must depart from the rational expectations benchmark of the model. To the best of our knowledge, the FFA market had never been examined from the perspective of a structural, behavioural economic model before.

C1683: **Pairs trading, technical analysis and data snooping: Mean reversion vs momentum**
Presenter: **Athanasios Pantelous**, Monash University, Australia
Co-authors: Ioannis Psaradellis, Jason Laws, Georgios Sermpinis
The aim is to examine technical trading rules performance of the statistical arbitrage investment strategy, pairs trading, using daily data over the period 1990-2016 for 15 commodity, equity and famous currency pairs. Adopting a previous false discovery rate test to control for data snooping bias and exercising 18,412 technical trading rules, we find evidence of significant predictability and excess profitability, especially for commodity spreads, where the best performing strategy generates an annualized mean excess return of 17.6%. In addition, we perform an out-of-sample analysis to cross-validate our results in different subperiods. We find that whilst the profitability of rules based on technical analysis exhibits a downward trend over the sample, the opportunities for pairs trading remains has increased in certain cases.

CO073 Room Jessel REGIME CHANGE MODELING II: MACRO AND POLICY Chair: Willi Semmler

C1648: **Credit cycles and inflation targeting in a regime switching model**
Presenter: **Willi Semmler**, New School for Social Research, United States
Co-authors: Timm Faulwasser, Marco Gross
A macro monetary policy model with double switching, one for the inflation and another one for the credit dynamics, is presented. We include in an Svensson's type inflation targeting model a nonlinear Phillips curve defining a state dependence of the inflation rate and add a dynamic equation for a state dependence of credit flows and credit spreads. As to the latter, we follow up the Minsky hypothesis that booms sow the seeds of the next crisis, in the sense that in credit expansions credit spreads are low and in contractions credit spreads are high. As to policies, we explore in a finite horizon monetary policy model the dynamic effects of price oriented as well as credit volume oriented monetary policy variants. This allows us to study stabilizing - destabilizing effects of price and non-price (credit volume) drivers of output gap, inflation and credit flows, in the context of our model. Empirically we are using a multi regime GVAR model to explore the double switching dynamics using Euro area data. We can demonstrate that in fact empirically, using EU data, both the output gap dynamics and the credit dynamics undergo regime switches. Moreover, the impulse-responses are explored using MRGVAR showing that the results are similar as predicted in the theoretical model.

C1726: **Financial fragmentation and the monetary transmission mechanism in the Euro area: A smooth transition VAR approach**
Presenter: **Ibrahim Tahri**, PIK (Potsdam Institute for Climate Impact Research), Germany
Co-authors: Hans-Helmut Kotz, Willi Semmler
The aim is to investigate the effect of financial fragmentation on the monetary transmission mechanism in the four largest euro area economies; Germany and France as core members and Italy and Spain as peripheral countries. We analyze the effects of financial fragmentation on the monetary transmission mechanism through the traditional interest rate channel. To analyze the impact of changes in policy rates on the behavior of real variables such as aggregate output and employment we employ a Smooth Transition VAR (VSTAR) model. Using a nonlinear multivariate time series model helps us capture the regime-dependent dynamics of the variables under study. The results obtained show that money market rates targeted by the central bank do not completely pass through to banks lending rates to firms, hence supporting the hypothesis of an impairment of the monetary transmission mechanism as a result of financial fragmentation.

C1655: **Sign restrictions in smooth transition VAR models**
Presenter: **Martin Bruns**, DIW Berlin and Free University Berlin, Germany
Co-authors: Michele Piffer
A way of identifying Smooth Transition Vector Autoregressive models using sign restrictions is developed. In so doing, we offer an alternative to the recursive identification typically used for these models. What makes sign restrictions tractable despite the nonlinearities of the model is the use of a Bayesian approach. We then use the model and argue that monetary policy shocks have a smaller effect on output when the economy is experiencing relatively high levels of uncertainty. The results cannot be driven by the higher price flexibility documented by the literature in periods when uncertainty is high. In fact, we find that not only output but also prices display a limited response to monetary shocks when uncertainty is high. The results suggest that the effectiveness of monetary policy is driven by mechanisms that go beyond price stickiness.

C1752: **Money and price developments in a historical timescale perspective: Implications for monetary analysis**
Presenter: **Marco Gallegati**, Polytechnic University of Marche, Italy
Co-authors: Michele Fratianni, Federico Giri
The aim is to deepen monetary analysis by providing greater insight into the association between trends in monetary growth and inflation using wavelet exploratory analysis. Several application of wavelet analysis to the money-inflation relationship have been recently provided, but no one of them has used a dataset covering a long time span, which is essential when examining a long-run time-varying relationship. A historical dataset developed previously allows the investigation of the relationship between money growth and inflation for several developed countries over the period 1870-2011, thus covering different monetary policy regimes, periods characterized by high and low inflation level and volatility, and periods of money demand instability. CWT tools such as the wavelet coherence and phase analysis allows to explore the money growth-inflation relationship within a unified framework where, for each frequency band, we can study whether co-movements in money growth and inflation have become stronger or weaker through time and whether there have been changes in the lead-lag patterns between the two variables. In particular, by detecting the specific frequency bands where the long-run link between money growth and inflation are most significant we can identify a measure of core or underlying monetary growth with the ability to enhance the operativeness of monetary developments for the conduct of monetary policy aiming at price stability.

C1751: **Convergence clubs in the European Union**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Harry Haupt, Willi Semmler
Growth convergence for countries of the European Union and the Euro-area is analyzed. Employing a recent approach to data-drivenly determine the club-membership of the countries, we find considerable variation in club composition over countries and time.

CO158 Room Montague MODELING AND FORECASTING IN FINANCIAL ECONOMETRICS Chair: Helena Veiga

C0882: **On testing financial contagion and modelling time-varying volatility interactions**
Presenter: **Cristina Amado**, University of Minho, Portugal
Volatility contagion is analyzed by testing significant increases in cross-market correlations of return volatilities. For modelling the dynamics of the volatility interactions, we propose an additive time-varying (or partially time-varying) structure in which a time-dependent component is added to the extended vector GARCH process. In this setting, the co-dependence in volatility is allowed to change smoothly between two extreme volatility regimes and contagion is identified from these crisis-contingent structural changes. A Lagrange multiplier type test of volatility contagion is proposed for testing the null hypothesis of constancy co-dependence against a smoothly time-varying interdependence. Finite sample properties of the proposed test statistic are investigated by Monte Carlo experiments.

C0889: **On conditional dynamic skewness and directional forecast of currency exchange rates**
Presenter: **Julien Hambuckers**, University of Gottingen, Germany
A GARCH-in-mean model is formulated where the innovations follow a non-Gaussian sinh-arcsinh distribution, with time varying skewness parameter. The structural equation of this parameter is assumed to be of the ARMA-X type, allowing for an effect of past stochastic shocks, autoregressive terms and interest rate differential on conditional skewness dynamic. This model is used to predict the direction of change of three major currency pairs (USD/EUR, USD/GBP and USD/CHF) over the period 1999-2016. We find a dynamic skewness specification to be highly significant and able to improve the correct classification rate between up and down periods. Robust bootstrap tests support these findings. Economically speaking, we show that, notwithstanding a better in-sample performance compared to benchmark models for all pairs, we are able to obtain a better and positive out-of-sample performance for the USD/CHF pair.

C1011: **Dynamic mixed frequency pooled copula**
Presenter: **Audrone Virbickaite**, University of Konstanz, Germany
Co-authors: Hedibert Lopes
Modeling dependence between financial returns through copula by exploiting the information available from daily and intraday data is proposed. The two alternative copula models, obtained from low and high frequency data, are combined via density pooling approach and show superior performance in terms of sequential predictive log Bayes factors. The intraday copula is estimated via two alternative specifications: hierarchical and mixture model. While hierarchical model is the best in class of models that are based only on high frequency data, incorporating low frequency information provides additional gains in predictive model performance.Proposed pooled model is applied to 3-variate and 10-variate daily log returns of financial assets, traded at the NYSE.

C1822: **Volatility forecasting gains from jumps: On the effect of the nature of the jumps**
Presenter: **Marwan Izzeldin**, Lancaster University Management School, United Kingdom
Co-authors: Rodrigo Hizmeri, Anthony Murphy, Mike Tsionas
A growing literature documents gains in forecasting return volatility using high frequency data. More recently, attention has shifted to the identification and use of jumps in forecasting. We partition jumps into finite and infinite variance (infrequent large vs frequency small) jumps, and use an extended heterogeneous, autoregressive (HAR) model to examine the forecasting gains at various horizons from this decomposition. Inter alia, we consider intraday periodicity, robust-to-noise volatility measures, different sampling frequencies and regime changes. The forecasting gains from using finite jumps exceed those from using infinite jumps except at high frequencies. Using robust-to-noise volatility measures results in better forecasts. Filtering periodicity adversely affects both types of jump measures, but affects finite jumps to a greater extent. Based on the Giacomini and White test, we also find that more significant differences in forecasting gains at higher frequencies. Finally, during the crisis period, decomposing jumps into finite and infinite components improves the forecasts.

C1009: **Variance risk premium: Implications for financial and economic predictability**
Presenter: **Helena Veiga**, Universidad Carlos III de Madrid and BRU-IUL (Instituto universitario de Lisboa), Spain
Co-authors: Isabel Casas, Xiuping Mao
Using the decomposition of squared VIX into conditional variance of the stock market and equity variance risk premium, we explore the predictive power of both variables for stock returns, economic activity and financial instability. Measures of conditional variance and variance risk premium are obtained from new parametric and nonparametric asymmetric extensions of the heterogeneous autoregressive model (HAR). We find, using the new specifications, that the equity variance risk premium is a predictor of future stock returns while conditional variance is often rejected as a predictor for short and moderate prediction horizons. Furthermore, both variables are also predictors of financial instability and for some horizons of the economic activity. All in all, new nonparametric asymmetric measures of realized variance improve, considerably, the predictive power of the variance risk premium when considering stock market returns and financial instability.

CO190 Room Senate ROBUSTNESS, CONTAGION, DEPENDENCE, EXTREMES AND HEAVY TAILS Chair: Rustam Ibragimov

C0527: **New robust inference for predictive regressions**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration, Russia
Co-authors: Rustam Ibragimov
New simple approaches are proposed to robust inference in predictive regressions. First, we utilize instrumental variable estimators such as Cauchy estimator to guarantee the asymptotic normality of the estimator regardless of order of integration of the variables in regression models and endogeneity. Second, we apply previous results that show that robust inference on unknown parameters of interest under heterogeneity and dependence may be conducted by partitioning the data into some number of groups and performing the standard t-test with the parameter's group estimates and critical values of Student-t distributions. The proposed approach compares favorably with widely used alternative inference procedures in terms of its finite sample properties and is robust to different forms of nonstationary volatility and heavy tails in terms of size control whereas its power is comparable to other tests.

C1054: **The determinants of heavy-tailedness of stock returns in the Russian market**
Presenter: **Andrei Ankudinov**, Innopolis University, Russia
Co-authors: Oleg Lebedev
Using robust approaches we analyzed the degree of heavy-tailedness of Russian companies stock returns, as well as positive and negative returns asymmetry. We also considered the factors affecting heavy-tailedness and returns asymmetry. According to the obtained results, the Russian stock market appeared to be somewhat more heavy-tailed when compared to the stock markets of developed as well as emerging economies. Tail indices for individual companies lie in the interval (1,4), first moments are finite for all shares, while second and third tiers are generally characterized by much larger extreme outliers and can have infinite variances. It seems that the stocks liquidity has the strongest effect on the heavy-tailedness of returns: more liquid companies have less heavy right and left tails. A similar, though less statistically reliable effect is produced by the company size. At the same time, no significant effects on tail indices of sectoral affiliation, cross-listing, adding into quotation lists, state ownership were revealed. As for the influence of the above-mentioned factors on the asymmetry of returns, the reliability of relevant models is rather low. Most probably, the inclusion of a company in the quotation list of a stock exchange is associated with a somewhat lower likelihood of returns asymmetry. There are also some indications of a higher likelihood of returns asymmetry for companies representing the financial sector of the economy.

C1119: **Robustness, market (non-)efficiency, volatility slustering, stock return predictability and beyond**
Presenter: **Rustam Ibragimov**, Imperial College London and Innopolis University, United Kingdom
The analysis of (in-)efficiency, volatility clustering and predictive regressions in financial markets using traditional approaches based on ACF's of squared returns and asymptotic methods is complicated by the presence of non-linear dependence (e.g., modelled using GARCH-type dynamics) and heavy-tailedness. Similar problems appear with commonly used used predictive regressors. Several new approaches are presented to deal with the above problems. The approaches are based on the use of autocorrelations of absolute values of the returns, together with new methods of robust inference using conservativeness of t-statistics. In the methods, estimates of parameters of interest are computed for groups of data and the inference is based on t-statistics in resulting group estimates. This results in valid robust inference under a wide rage of heterogeneity and dependence assumptions under the only conditions of asymptotic normality of group estimates. Numerical results and empirical applications confirm advantages of the new approaches over existing ones and their wide applicability in the study of market (in-)efficiency, volatility clustering, predictive regressions and beyond.

C0656: **Extreme returns in the Russian stock market: Unexpected tails in risk measures**
Presenter: **Oleg Lebedev**, Innopolis University, Russia
Co-authors: Andrei Ankudinov
The purpose is to investigate extreme fluctuations in the Russian financial market using extreme value theory methods. For that, we: (i) determine the probability distribution that most accurately describes the behavior of extreme returns in the Russian stock market and further present the

analysis of the dynamics of its parameters over time; (ii) test the reliability of dynamic risk-management models based on the fitted distributions. A highly volatile Russian market provides a natural environment for the analysis of distributional properties of extreme changes in financial returns. At the same time, the high degree of heavy-tailedness and heterogeneity of observations dealt with require applications of robust statistical methods. The results show that the extreme returns parameters lie between the less heavy-tailed GEV distribution and the more heavy-tailed GLO distribution. However, when we move deeper into the tails of distributions, it is the GLO distribution that appears to be more suitable since other distributions fail to accurately estimate the probability of extreme returns. Accordingly, the GLO distribution appears to be more adequate for integration into risk-management models for the Russian stock market. Applications of less heavy-tailed distributions would result in systematic underestimation of the largest market slumps.

C0960: **A new measure of vector dependence, with applications to financial risk and contagion**
Presenter: **Artem Prokhorov**, Innopolis University, Australia
Co-authors: Ivan Medovikov

A new nonparametric measure of association between an arbitrary number of random vectors is proposed. The measure is based on the empirical copula process for the multivariate marginals, corresponding to the vectors, and is robust to the within-vector dependence. It is confined to the $[0, 1]$ interval and covers the entire range of dependence from vector independence to a monotone relationship element-wise. We study the properties of the new measure under several well-known copulas and provide a nonparametric estimator of the measure, along with its asymptotic theory, under fairly general assumptions. To illustrate the applicability of the new measure, we use it in applications to financial contagion, systemic risk, and portfolio choice. Specifically, we test for contagion effects between equity markets in North and South America, Europe and Asia, surrounding the financial crisis of 2008 and find strong evidence of previously unknown contagion patterns. In the context of sovereign bonds and credit default swaps, we study the evolution of systemic risk in European financial markets and uncover large differences from previous estimates. Finally, based on a real-time portfolio utilizing the new systemic risk estimates, we illustrate the implications of the new measure for portfolio choice and risk management.

CO738 Room Woburn RISKS AND FINTECH Chair: Dominique Guegan

C0771: **Forecasting inflection points: Hybrid methods with machine learning algorithms**
Presenter: **Julien Chevallier**, IPAG Business School, France
Co-authors: Bangzhu Zhu, Shujiao Ma, Yi Ming Wei

Hybrid time-series forecasting models are investigated which are based on combinations of ensemble empirical mode decomposition (EEMD) and least square support vector machines (LSSVM). Several algorithms are considered: genetic algorithm (GA); grid-search (GS); particle swarm optimization (PSO) and uniform design (UD). Theoretical guarantees of prediction accuracy are tested with sine curves. From a numerical testing perspective, we are interested to show the superiority of one approach over another based on multiple arbitrary time series in engineering (HTTP requests to the NASA servers), finance (MCD ticker on the NYSE), macroeconomics (China GDP) and commodities (ECX CO2 Futures). Out-of-sample forecasting superiority is assessed in a horse race based on Diebold-Mariano test statistics.

C1252: **Bitcoins and challenges for financial regulation**
Presenter: **Dominique Guegan**, Universite Paris 1 - Pantheon-Sorbonne, France

Limits and interest of the new crytocurrency, as well as questions relative to the financial regulation, are discussed.

C1185: **Impact of multimodality of distributions on VaR and ES calculations**
Presenter: **Kehan Li**, University Paris I Pantheon Sorbonne, France
Co-authors: Dominique Guegan, Bertrand Hassani

Unimodal probability distribution has been widely used for Value-at-Risk (VaR) computation by investors, risk managers and regulators. However, financial data may be characterized by distributions having more than one modes. Using a unimodal distribution may lead to bias for risk measure computation. We discuss the influence of using multimodal distributions on VaR and Expected Shortfall (ES) calculation. Two multimodal distribution families are considered: Cobb's family and distortion family. We provide two ways to compute the VaR and the ES for them: an adapted rejection sampling technique for Cobb's family and an inversion approach for distortion family. For empirical study, two data sets are considered: a daily data set concerning operational risk and a three month scenario of market portfolio return built with five minutes intraday data. With a complete spectrum of confidence levels from 0.001 to 0.999, we analyze the VaR and the ES to see the interest of using multimodal distribution instead of unimodal distribution.

C1539: **Regulatory learning: How to supervise machine learning models with an application to credit scoring**
Presenter: **Bertrand Hassani**, Universita Paris 1 Pantheon Sorbonne - Labex Refi, France
Co-authors: Dominique Guegan

The arrival of big data strategies is threatening the latest trends in financial regulation related to the simplification of models and the enhancement of the comparability of approaches chosen by financial institutions. Indeed, the intrinsic dynamic philosophy of Big Data strategies is almost incompatible with the current legal and regulatory framework, as illustrated. Besides, as presented in our application to credit scoring, the model selection may also evolve dynamically forcing both practitioners and regulators to develop libraries of models, strategies allowing to switch from one to the other as well as supervising approaches allowing financial institutions to innovate in a risk mitigated environment. The purpose is therefore to analyse the issues related to the Big Data environment and in particular to machine learning models highlighting the issues present in the current framework confronting the data flows, the model selection process and the necessity to generate appropriate outcomes.

C1713: **Blockchain towards legal recognition in the major economic countries**
Presenter: **Stephane Blemus**, Paris Sorbonne University, France

In just a few years, the blockchain has become a major topic for regulators. As this disruptive and decentralized technology has emerged as a key business topic for start-ups and market participants, the central banks and financial regulators have changed from an initial strong hostility to a more cautious and market-friendly position, notably in the United States and in the European Union. The aim is to extensively cover and compare the current regulatory trends in some relevant financial centres on the various applications enabled or issues raised by the blockchain technology (Bitcoin/cryptocurrencies, smart contracts, DAO or decentralized autonomous organization, ICO or initial coin offering).

CO104 Room SH349 VOLATILITY MODELS AND THEIR APPLICATIONS Chair: Yasuhiro Omori

C0811: **Particle rolling MCMC with forward and backward block sampling: Conditional sequential Monte Carlo update approach**
Presenter: **Naoki Awaya**, University of Tokyo, Japan
Co-authors: Yasuhiro Omori
The objective is to provide a new simulation-based methodology for rolling estimation in state space model from Bayesian approach. This type of estimation requires sampling by simulation-based method from a lot of posteriors if the model does not have so simple form. Repetition of sampling from each posterior by Markov Chain Monte Carlo is not realistic from a viewpoint of computational time so to address this problem a new sampling algorithm based on sequential Monte Carlo is presented. This method is applied to SP 500 data with the realized stochastic volatility with leverage model and how the economic structure which generates the financial data is changed is shown.

C0854: **Bayesian analysis of intraday stochastic volatility models with leverage and skew heavy-tailed error**
Presenter: **Teruo Nakatsuma**, Keio University, Japan
Co-authors: Makoto Nakakita
Intraday high frequency data of asset returns exhibit not only typical characteristics (e.g., volatility clustering), but also a cyclical pattern of return volatility that is known as intraday seasonality. We extend the stochastic volatility (SV) model for application with such intraday high frequency data and develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm for Bayesian inference of the proposed model. Our modeling strategy is two-fold. First, we model the intraday seasonality of return volatility with a B-spline polynomial and estimate it along with the stochastic volatility simultaneously. Second, we incorporate a possibly skew and heavy-tailed error distribution into the standard SV model by assuming that the error distribution belongs to a family of generalized hyperbolic (GH) distribution such as Laplace, variance-gamma and Student's t. As a demonstration of our new method, we estimate the proposed intraday SV model with 5-minute return data of gold futures in the Tokyo Commodity Exchange, and conduct a selection of model specifications with the deviance information criterion (DIC). The result shows that the SV model with Laplace error is the best, but it does not support the asymmetry of the return distribution.

C0991: **Predictability of excess bond premium and variance risk premium for business cycles and recession risk**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Masato Ubukata, Yoichi Ueno
Variance risk premium (VRP) is the difference between the conditional variance of stock returns under the risk-neutral measure and that under the physical measure. Empirical evidence suggests that the VRPs of stock indexes may predict some financial and macroeconomic variables such as the excess returns of stock indexes, credit spreads and business cycles. Excess bond premium (EBP) is a component of credit spreads that is not attributable to expected default risk. It has also attracted the attention as a predictor for business cycles and recession risk. Using the Japans data, the predictive power of VRP for business cycles and recession risk is compared with that of EBP. The Japanese volatility index (VXJ) published by Osaka University is used as the conditional variance of Nikkei 225 stock index under the risk-neutral measure. The conditional variance of Nikkei 225 under the physical measure is calculated using the extended heterogeneous autoregressive (HAR) model with the realized volatility of Nikkei 225. EBP is calculated using the credit spreads of corporate bonds in Japan.

C0962: **Factor multivariate realized stochastic volatility model**
Presenter: **Yuta Yamauchi**, University of Tokyo, Japan
Co-authors: Yasuhiro Omori
Although modelling time-varying volatility and correlations of multivariate asset returns is one of most important problems in the financial risk management, it has been difficult to obtain stable inference of covariance of asset returns due to the high dimensionality of parameters in dynamic covariance structure. One major solution to reduce the number of parameters is to introduce factor structure assuming that a small number of common factors describe the dynamics of time-varying covariance matrices as discussed in the factor stochastic volatility models. We propose parsimonious modelling of multivariate asset returns based on dynamic factor stochastic volatility models with leverage effect, which allow dynamic latent factors. Firstly, to stabilize the estimation and prediction of time-varying parameters, we incorporate additional observations based on intraday asset returns and market indices. We use realized measures for covariance of asset returns based on intraday asset returns such as realized covariances, and latent factors of asset returns such as market indices. Secondly, we reduce the number of parameters of leverage effect, omitting the leverage effect between each asset and each volatility of asset. We only introduce leverage effect between each latent factor and each volatility of asset.

C1762: **Realized stochastic volatility with skew-t error**
Presenter: **Makoto Takahashi**, Osaka University, Japan
Co-authors: Yasuhiro Omori, Toshiaki Watanabe
The realized stochastic volatility model with Student's t error is extended to the skew-t error model by using the mixture representation of the skew-t distribution. This representation, including normal, Student's t, and skew-normal distributions as special cases, allows flexible modeling of the skewness and heavy tails in the conditional distribution of financial asset returns. The Bayesian estimation scheme via a Markov chain Monte Carlo method is developed and the empirical results using the US and Japanese stock indices will be presented.

CC707 Room G5 CONTRIBUTIONS IN TIME SERIES ANALYSIS Chair: Richard Luger

C1478: **Explosive asset price bubble detection with unknown bubble length and initial condition**
Presenter: **Emily Whitehouse**, University of Nottingham, United Kingdom
Recent research has proposed a method of detecting explosive processes that is based on forward recursions of OLS, right-tailed, Dickey-Fuller [DF] unit root tests. An alternative approach using GLS DF tests is considered. We derive limiting distributions for both mean-invariant and trend-invariant versions of OLS and GLS-based PWY test statistics under a temporary, locally explosive alternative. These limits are shown to be dependent on both the value of the initial condition and the start and end points of the temporary explosive regime. Local asymptotic power simulations show that a GLS version of the PWY statistic offers superior power when a large proportion of the data is explosive, but that the OLS approach is preferred for explosive periods of short duration as a proportion of the total sample. These power differences are magnified by the presence of an asymptotically non-negligible initial condition. We propose a union of rejections procedure that capitalises on the respective power advantages of both OLS and GLS-based approaches. This procedure achieves power close to the effective envelope provided by the two individual PWY tests across all settings of the initial condition and length of the considered explosive period. These results are shown to be robust to the point in the sample at which the temporary explosive regime occurs.

C1640: **The impact of moving holidays on official statistics time series**
Presenter: **Bethan Russ**, Office for National Statistics, United Kingdom
Co-authors: Tariq Aziz
The Office for National Statistics (ONS) is the UKs largest independent producer of official statistical and the recognised national statistical institute

of the UK. The Time Series Analysis Branch (TSAB) within the ONS is responsible for the seasonal adjustment and analysis of tens of thousands of economic and social ONS time series. A major challenge faced when seasonally adjusting time series is accounting for annual events that move dates from one calendar year to the next, for example, Easter. If these events are not accounted for appropriately it will impact on the estimation of seasonal factors, and leave systematic calendar related effects in the seasonally adjusted series. Currently TSAB tests for Easter effects and, if identified, estimates and removes them as part of seasonal adjustment. This method assumes that daily activity changes by a fixed amount or proportion for a given number of days before Easter Sunday and remains at this level until Easter Saturday. There are other moving holidays celebrated in the UK, which may have an impact on time series despite not being public holidays. These are Chinese New Year, Ramadan, Eid al-Fitr and Eid al-Adha. Currently these holidays are not adjusted for in any seasonally adjusted time series published by the ONS. TSAB has undertaken research to test alternative windows for Easter effects and whether other moving holidays have identifiable effects on ONS time series.

C1585: **Maximum likelihood estimation for noninvertible ARMA processes with stable distribution innovation process**
Presenter: **Juho Nyholm**, University of Helsinki, Finland

The focus is on properties of a maximum likelihood estimators of the nonivertible ARMA process with innovation process in the class of stable distributions. We restrict the analysis to non-Gaussian stable distributions that do not necessarily have finite second moments. We show that there is a non-standard distribution for the sequence of $n^{1/\alpha}$-consistent estimators that are obtained as maximizers of the likelihood function. The non-standard limiting distribution of the ML estimator is obtained as a maximizer of a random function. Results are applied to economic time series data and we show that relaxing the standard assumptions on Gaussianity and invertibility may lead to better fit and alter the conclusions drawn from the data.

C1451: **Resampling uncertainty of principal components factors**
Presenter: **Javier de Vicente Maldonado**, Carlos III University, Spain
Co-authors: Esther Ruiz

In the context of Dynamic Factor Models (DFMs), one of the most popular procedures for factor extraction is Principal Components (PC). Measuring the uncertainty associated to PC factor estimates should be part of interpreting them. Although the asymptotic distribution of PC factors is known, it could not be an appropriate approximation to the finite sample one for the sample sizes and cross-sectional dimensions usually encountered in practice. The main problem is that it does not take into account parameter uncertainty. Alternatively, several bootstrap procedures have been proposed in DFM with goals related to inference. We show that these procedures are not appropriate to measure the uncertainty of PC factor estimates and propose an alternative resampling procedure designed with this purpose. The asymptotic and finite sample properties of the proposed procedure are analyzed and compared with those of the asymptotic and alternative extant bootstrap procedures. The results are empirically illustrated obtaining confidence intervals of the underlying factor in a system of Spanish macroeconomic variables.

C1792: **An evaluation of automatic outlier detection methods**
Presenter: **Jennifer Davies**, Office for National Statistics, United Kingdom

The Office for National Statistics (ONS) is the UKs largest independent producer of official statistical and the recognised national statistical institute of the UK. The Time Series Analysis Branch (TSAB) within the ONS is responsible for the seasonal adjustment and analysis of tens of thousands of time series measuring economic and social phenomena. A major challenge faced when seasonally adjusting time series is identifying and accounting for outliers; observations within a time series that differ distinctly from the general pattern of the trend and/or seasonal components. An outlier, if unidentified or misspecified, may cause issues in the analysis of a time series and lead to large revisions. Revisions are an inevitable part of the production process due to late returns and the process of seasonal adjustment. Analysts aim to minimise revisions to ensure reliability and retain public faith in published estimates, hence the importance in correctly identifying and specifying outliers. Outliers at the end of a time series can be particularly difficult to identify in a timely manner, due to the uncertainty around future observations.Techniques to account for outliers at the end of a series will be evaluated. It will consider the currently used automatic outlier detection method in X13-ARIMA-SEATS and compare this with other outlier detection methods found in the literature. The main alternative methods considered are indicator saturation and changepoint methods.

CG074 Room Court CONTRIBUTIONS IN SPATIAL ECONOMETRICS Chair: Arnab Bhattacharjee

E1617: **Spatial dependence and unobserved heterogeneity in stochastic frontier models: A Bayesian approach**
Presenter: **Antonio Carvalho**, Heriot-Watt University, United Kingdom

The aim is to contribute to the literature of panel data Stochastic Frontier modelling and efficiency measurement in production and cost functions by discussing estimation of efficiency and other parameters of interest in the context of spatial dependence and unobserved heterogeneity. The proposed Bayesian approach allows for the decomposition of efficiency into a time-varying component and a persistent component in a random effects model with spatial spillovers. The measurement of efficiency with spatial spillovers is more complex than in the non-spatial case and measurement methods are discussed in relative and absolute scales. A Bayesian approach with a standard assumption of half-normal distribution for efficiency is outlined. Small sample performance of the model is deeply related to the underlying signal-to-noise ratios with good performance for larger samples and encouraging results for applied researchers. The approach is applied to aggregate productivity in European countries between 1992 and 2005, with a particular focus on transition economies and the role of spatial dependence and unobserved heterogeneity in the production frontier. Results show a large amount of persistent inefficiency which would be ignored under less complex estimation methods, and also a non-negligible amount of spatial dependence.

C1714: **Saddlepoint techniques for spatial panel data models**
Presenter: **Chaonan Jiang**, University of Geneva, Switzerland
Co-authors: Davide La Vecchia, Elvezio Ronchetti, Olivier Scaillet

New higher-order asymptotic techniques are developed for the Gaussian maximum likelihood estimator (henceforth, MLE) of the parameters in a spatial panel data model, with fixed effects, time-varying covariates, and spatially correlated error. The first-order asymptotics needs the cross-sectional sample size (n) to diverge, while the time dimension (T) can be fix. The resulting approximation to the MLE density has absolute error of order $O(m^{-1/2})$, for $m = n(T-1)$. We illustrate that, when n and T are small, the first-order asymptotics can be inaccurate, specially in the tails – the parts of the density we are typically interested in, e.g. for the p-values. To improve on the accuracy of the extant asymptotics, we introduce a new saddlepoint density approximation, which features relative error of order $O(m^{-1})$. The main theoretical tool is the tilted-Edgeworth technique, which, by design, yields a density approximation that is always non-negative and does not need resampling. We provide an algorithm to implement our saddlepoint approximation and we illustrate the good performance of our method via numerical examples. Monte Carlo experiments show that, for the spatial panel data model with fixed effects and $T = 2$, the saddlepoint approximation yields accuracy improvements over the routinely applied first-order asymptotics and Edgeworth expansions, in small to moderate sample sizes, while preserving analytical tractability.

C0245: **Determinants of price differences in Russian regions**
Presenter: **Yury Perevyshin**, The Russian Presidential Academy of National Economy and Public Administration, Russia

Based on econometric methods for analyzing panel data, factors that explained price differences in the 76 Russian regions in the period 2000-2015 were determined. An explanation of the mechanisms of their influence was proposed. Problems of different price levels in countries or regions using a common currency can were listed. We tested fixed time effects model vs. pool model, took into consideration spatial autocorrelation between regions. The results show that a steady effect on regional price differences both during the whole period of the study and at individual subperiods was provided by:a) the Balassa-Samuelson effect, b) costs of trade, c) the mismatch of business cycle phases, d) the structure of the regional economy, e) the regional structure of household incomes. Regional income and price level are subject to spatial autocorrelation. The results obtained can be applied to the development of measures of regional economic policy aimed at synchronization of inflation processes, while forecasting both national and regional inflation, as well as modeling the consequences of the single monetary policy of the Bank of Russia for regional economies.

C0646: **Semiparametric spatio-temporal energy data analysis**
Presenter: **Xiaohang Ren**, University of Southampton, United Kingdom
Co-authors: Zudi Lu

Energy plays a critically important role in the development of the economy and there are a large number of spatiotemporal data with complex structures in energy research. We investigate the relationship between natural gas prices and oil prices at US state-level by constructing a novel semiparametric spatiotemporal model to estimate the impacts of oil price shocks on the US natural gas prices in different states. The proposed estimators are studied with numerical simulations carried out for comparisons of the properties of the estimates. This approach can capture the nonlinear dependence between the natural gas prices and oil price in the US states and uncover price shock transmission and nonlinear cross-market dependency of the natural gas markets.

C1661: **Spatio-tempral autoregressive conditional heteroskedasticity model**
Presenter: **Takaki Sato**, Tohoku University, Japan
Co-authors: Yasumasa Matsuda

A spatio-temporal extension of time series autoregressive conditional heteroskedasticity (ARCH) models is proposed. We call the spatio-temporally extended ARCH models as spatio-temporal ARCH (ST-ARCH) models. ST-ARCH models specify conditional variances given same time surrounding observations and past time observations, which constitutes a good contrast with time series ARCH models that specify conditional variances given past observations. A spatial weight matrix which quantify the closeness between observations is used to express effects from surrounding observations. We estimate the parameters of ST-ARCH models by a two-step procedure of least squares and the quasi maximum likelihood estimation to avoid bias of least squares estimators. We demonstrate the empirical properties by real data analysis of stock price data in the Japanese market to show the relation between volatility of a particular stock and change rates of same time and past time other stock prices.

CG004 Room G3 CONTRIBUTIONS IN CREDIT AND SYSTEMIC RISK Chair: Genevieve Gauthier

C1350: **Nonlinear decomposition banks credit ratings**
Presenter: **Patrycja Chodnicka - Jaworska**, University of Warsaw, Poland

The purpose is to analyze the impact of the business cycle using non-linear decomposition of credit ratings on banks notes. A literature review was made and the following research hypothesis has been put: The banks credit ratings are pro-cyclical. There were used ordered logit panel data models, and as a dependent variable there were taken the long-term issuer credit ratings published by the three largest rating agencies, Fitch, S&P and Moody. There have been collected data from the Thomson Reuters Database. The analysis was based on decomposition method that uses CDS spreads. The data were collected for the years 1990 - 2016. As an independent variables there have been used CAMEL factors. Data was obtained from the World Bank and Thomson Retuters database.

C0775: **Market credit risk in Europe**
Presenter: **Ana-Maria Dumitru**, University of Surrey, United Kingdom
Co-authors: Tom Holden

The run-up to the Greek default featured marked increases in the cost of insuring sovereign debt from almost all European countries, as evidenced by their credit default swap (CDS) rates. One explanation for the perceived higher default risk in non-periphery countries is that market participants believed a default in the periphery might increase the risk of a future default in the core. To test for such a dynamic contagion between credit related events in differing countries, we develop a procedure for tractably estimating high-dimensional Hawkes models using CDS prices. We escape the curse of dimensionality thanks to modelling the market portfolio of risk across countries, which serves as a sort of common factor. We further reduce dimensionality by taking a maximum-likelihood approach to estimation, avoiding having event intensities in the parameter set. Our approximation to the likelihood converges to the true likelihood as the sampling frequency goes to infinity; this ensures consistency even given non-stationarity, and permits tight identification. We find little evidence of shocks to one European country having an instantaneous effect on others, but risk in one country does gradually pull up risk in others.

C1523: **Credit risk model applied to the agricultural sector**
Presenter: **Maria Rosa Nieto Delfin**, Investigaciones y Estudios Superiores SC, Mexico

Credit risk management has been widely studied especially when it comes to managing the risk that banks face for default in payment of credit cards. With regard to the agricultural sector in Latin America, management of default risk has not been analyzed properly. The aim is to apply the CyRCE model developed by the Mexican Central Bank, and so far applied solely to measure credit risk in commercial banks. According to our knowledge is the first time this model is applied in the agricultural sector for Mexico. Some modifications have been made to the original model and it has been empirically applied to a client portfolio of a Peasant Financial Company. The results show risk measures according with those expected by risk managers.

C1691: **Forecasting systemic risk**
Presenter: **Juan-Angel Jimenez-Martin**, Complutense of Madrid, Spain
Co-authors: Massimiliano Caporin, Laura Garcia-Jorcano

The Conditional Value at Risk (CoVaR), the VaR of the financial system conditional on an institution being in financial distress, has been previously modified by proposing a change in the definition: from the maximum loss of the system conditional on the financial institution being in its VaR, to the financial institution being at most at its VaR. We extend this methodology. The first objective is to evaluate whether the multivariate GARCH specification can be relevant for forecasting CoVaR. Additionally to the DCC model used in GE, we use the BEKK model and the Orthogonal GARCH (OGARCH) model. The second novelty is about the distributional assumption on the error. We use Filtered Historical Simulation (FHS) that has emerged as one of the best tools for calculating risk measures as VaR and consequently a robust alternative as a procedure to forecast CoVaR. The third contribution is to produce 1-day and 10-day ahead CoVaR forecasts. Finally, in order to assess the validity of the systemic risk

measure forecast we run, not only the traditional statistical based unconditional and conditional coverage tests, but we also use a variety of loss function applied by banks and regulators to evaluate VaR model performances.

C1592: **The impact of mandatory amortization of mortgage loans on the housing market**
Presenter: **Aleksandar Petreski**, Jonkoping International Business School, Sweden
Co-authors: Andreas Stephan, Urban Osterlund
Using transactions data on apartment sales in Sweden, and taking in consideration the decision made by Swedish Riksbank, we examined the effect of mandatory mortgage amortization on apartment prices. In addition, we examined the effect on apartment demand preferences. For this purpose, we created indices of apartment prices using traditional hedonic models. We applied the spatial model in estimating a zone level apartment price index. With the created indices, we studied the effect on apartment prices by using the traditional panel data method. The propensity score model was used to analyze the effect of the decision on the property prices. We applied different time periods as the event window (quarter, 2 quarters, 1 year, 2 years). Finally, we tested several cut-off time points. We found no effect of the Riksbank decision on apartment prices, however we found a change in apartment preferences.

CG159 Room G4 CONTRIBUTIONS IN RETURN PREDICTABILITY Chair: Cees Diks

C1584: **Directional predictability of daily stock returns**
Presenter: **Janis Becker**, Leibniz University Hanover, Germany
Co-authors: Christian Leschinski
In contrast to the monthly or quarterly case, daily stock returns are generally regarded as unpredictable. While this may be true for the level of daily returns, we focus on the signs of these returns. Using various machine learning techniques, such as neural networks and general additive models, we show that meaningful directional forecasts can be generated. The analysis is carried out using pseudo out-of-sample forecasts for a data set consisting of all stocks in the Dow Jones Industrial Average from 2003-2016. Relevant predictor variables are chosen beforehand - in a separate model selection window from 1996-2003. This model selection procedure is carried out using a cross-validation procedure with forward chaining that is applicable in a time series context. Since the forecast period and the model selection period are strictly separated, the procedure mimics the situation a forecaster would face in real time. Applying common statistical tests, our forecasts are shown to be statistically significant. Therefore, we draw the conclusion that the sign of daily stock returns is (to some extent) predictable. Moreover, trading strategies based on these forecasts suggest the possibility to outperform the market index in terms of return and Sharpe ratio.

C1700: **Predicting intraday returns based on overnight returns for the US stock market**
Presenter: **Hao Li**, University of Amsterdam, Netherlands
Co-authors: Cees Diks, Valentyn Panchenko
Research on high frequency data for the S&P 500 Exchange-traded Fund (ETF) from 2003 to 2013, using Trade and Quote (TAQ) data, documents nonlinear dependence between overnight returns and subsequent intraday returns based on both parametric and nonparametric models. Specifically, the dependence between overnight returns and the first half-hour returns is significantly negative while the dependence between overnight returns and the last half-hour returns is significantly positive. However, although the returns in-between the first half-hour and the last half-hour do not linearly depend on overnight returns, using parametric as well as nonparametric methods we find them to depend on overnight returns nonlinearly.

C1638: **How learning from macroeconomic experiences shapes the term structure of interest rates**
Presenter: **Kasper Jorgensen**, Aarhus University, Denmark
Constant-gain learning expectations of consumption growth and inflation capture 81-88 pct of the variation in 1- through 10-year nominal interest rates. The residuals in interest rates from the learning-based factors have a two-factor structure and represent investor sentiment. The four term structure factors predict excess returns with R2's up to 53 pct, and subsume the predictive information in the most popular bond return predictors. We use these four factors to formulate a dynamic term structure model, which implies that short rate expectations account for the trend component of long term interest rates, whereas term premia are more cyclical, compared to the predictions of the workhorse affine term structure model. We also show that the proposed model is consistent with a declining equilibrium real rate over the last two decades.

C1630: **Common cycles in volatility and cross section of stock returns**
Presenter: **Lucie Kraicova**, Charles University, Faculty of Social Sciences, Czech Republic
Co-authors: Jozef Barunik
The purpose is to study the relationship between conditional quantiles of returns and the long-, medium- and short-term volatility in a portfolio of financial assets. We argue that the combination of quantile panel regression and wavelet decomposition of the volatility time series provides us with new insights into the pricing of risk and increases the accuracy of our estimates of return quantiles. Our results contribute to the literature on the risk-return relationship with an emphasis on portfolio management under various investment horizons. Moreover, the analytical framework that we introduce should be applicable to a wide range of problems outside of our research area.

C0289: **On the usefulness of financial variables to predict the conditional distribution of the market portfolio**
Presenter: **Azam Shamsi Zamenjani**, McMaster University, Canada
The predictability of the conditional distribution of market returns using financial and macroeconomic variables is investigated. In contrast to the extant literature that mainly focuses on the mean forecasts, we study the predictability of the entire density. This provides us with useful information on the uncertainty around the point forecasts and tail events that is valuable in areas such as asset allocation and risk management. We consider a Bayesian nonparametric mixture model that allows the mixing distribution to change with time. The weights of the mixture are constructed as Probit transformations of a linear combination of the predictors. This allows us to study whether these predictors are useful in forecasting the unknown and time-varying density of the innovations. We compare statistical and economic measures of forecasting performance of the proposed model with a set of benchmark models. In spite of little or no improvement in point forecasts, certain variables display significant out-of-sample predictive ability for the stock return density, and increase economic value for investors when employed in portfolio decisions.

CG076 Room Gordon CONTRIBUTIONS IN VECTOR AUTOREGRESSION Chair: Ralf Brueggemann

C1538: **Residual bootstrap for VAR models estimated by least absolute deviations**
Presenter: **Hanno Reuvers**, Maastricht University, Netherlands

The fixed design residual bootstrap method is considered in order to conduct inference in stationary vector autoregressive models. The fixed design residual bootstrap treats the regressor matrix as fixed and adds resampled residuals to construct a bootstrap sample. The asymptotic validity of the method is established. We also show that bootstrapped Wald and LM type of test statistics can be used to test linear hypothesis. Our method does not rely on density estimation and is thus easy to apply. A Monte Carlo study reports on the finite sample performance. The fixed design bootstrap is suggested to have good size properties and better size properties than a recursive bootstrap scheme. The bootstrapped version of the Wald type of test has a higher power than both the bootstrapped version of the LM test and the asymptotically pivotal tests that require density estimation. The bootstrap Wald type of test is therefore recommended to practitioners.

C1761: **Reduced rank regression in a large VAR model**
Presenter: **Kelly Trinh**, University of Queensland, Australia
Co-authors: Rodney Strachan

The aim is to address the over-parameterization issue in large vector autoregressive (VAR) models by using reduced rank regression. A model specification will be provided which is invariant to the variable ordering. To select the rank of VAR coefficients, we consider a marginal likelihood approach which is approximated by cross entropy, predictive likelihood, and Laplace approximation. We carry out an extensive Monte Carlo simulation to examine the performance of these approaches in rank selection. The results suggest that these approaches underestimate the rank of VAR coefficients when the dimensions of VAR systems grow, and when the singular values of the VAR matrices are small (close to zero). We go further to examine the forecast performance of the misspecified rank models relative to the model with the correct rank (the benchmark model) using the measures of point forecast (e.g., mean squared forecast error, weighted mean squared forecast error) and density forecast (e.g., log predictive likelihood). Our results suggest that the models with lower rank perform worse than the benchmark for short forecast horizons, however, they perform as well as or even beat the benchmark for long forecast horizons. These patterns are more evident when the magnitudes of the singular values of the VAR coefficient are small.

C1431: **Directed graphs and variable selection in large vector autoregressive models**
Presenter: **Ralf Brueggemann**, University of Konstanz, Germany
Co-authors: Christian Kascha

The dynamic relation among variables in vector autoregressive (VAR) models is represented as directed graphs. Based on these graphs, we identify so-called strongly connected components (SCCs). Using this graphical representation, we consider the problem of variable selection. We use the relations among the strongly connected components to select variables that need to be included in a VAR if interest is in forecasting or impulse response analysis of a given set of variables. We show that the set of selected variables from the graphical method coincides with the set of variables that is multi-step causal for the variables of interest by relating the paths in the graph to the coefficients of the 'direct' VAR representation. Empirical applications illustrate the usefulness of the suggested approach: Including the selected variables into a small US monetary VAR is useful for impulse response analysis as it avoids the well-known 'price-puzzle'. We also find that including the selected variables into VARs typically improves forecasting accuracy at short horizons.

C1634: **A mixed-frequency Bayesian vector autoregression with a steady-state prior**
Presenter: **Sebastian Ankargren**, Uppsala University, Sweden
Co-authors: Yukai Yang, Mans Unosson

A Bayesian vector autoregressive (VAR) model is proposed which allows for the explicit specification of a prior for the variables' "steady states"(unconditional means) for data measured at different frequencies, without the need to aggregate data to the lowest common frequency. Using a normal prior for the steady state and a normal-inverse Wishart prior for the dynamics and error covariance, a Gibbs sampler is proposed to sample from the posterior distribution. Moreover, a numerical algorithm for computing the marginal data density is suggested, which can be used to find appropriate values for the necessary hyperparameters. The proposed model is evaluated in an empirical forecasting situation in which Swedish GDP growth is being forecasted. The results generally indicate that the inclusion of monthly data is helpful and improves the accuracy of quarterly forecasts, in particular for shorter horizons.

C1590: **Dynamics of foreign direct investment in ASEAN-5: A vector autoregressive analysis**
Presenter: **Rutcher Lacaza**, University of the Philippines, Philippines
Co-authors: Stephen Jun Villejo

A vector autoregressive (VAR) model was used to examine movements of foreign direct investment (FDI) of the ASEAN-5 countries (Indonesia, Malaysia, Philippines, Singapore, and Thailand) in relation to other macroeconomic variables particularly interest rate, inflation rate, exchange rate, and trade openness. There is an interrelationship in the dynamics of FDI among countries in the region. For the Philippines, interest rate and inflation rate has an immediate positive effect on foreign direct investment. FDI of Philippines has a contemporaneous substitution effect with the FDI of Thailand and a lagged complementary effect with Singapore. For Malaysia, the model is the same with that of Philippines although the significant coefficient of inflation rate is on the second lag. Both the FDI of Philippines and Singapore have a complementary effect with the FDI of Malaysia. For Thailand, trade openness and exchange rate were significant in the model. FDI of Thailand has a complementary effect with the FDI of Malaysia. Shocks on the significant macroeconomic variables have a significant effect on FDI at certain periods. The variance decomposition of all models shows that a significant proportion of the movements in FDI is due to shocks on the other variables especially at very short horizons.

Sunday 17.12.2017 08:40 - 10:20 Parallel Session G – CFE-CMStatistics

EI011 Room Beveridge Hall TIME SERIES AND SPATIAL DATA: SCALABILITY AND APPLICATIONS Chair: Michele Guindani

E0161: **High-dimensional Bayesian geostatistics on modest computing environments**
Presenter: **Sudipto Banerjee**, UCLA, United States

With the growing capabilities of GIS databases, statisticians today routinely encounter geographically referenced data containing observations from a large number of spatial locations and time points. Over the last decade, hierarchical spatio-temporal process models have become widely deployed statistical tools for researchers to better understand the complex nature of spatial and temporal variability. However, fitting such models is often prohibitive in terms of storage and floating point operations with complexity increasing in cubic order for the number of spatial locations and temporal points. This renders such models unfeasible for large data sets. The aim is to discuss model-based strategies to circumvent computational bottlenecks using well-defined massively scalable spatio-temporal stochastic processes. Full Bayesian inference is sought and achieved for millions of locations on very modest computing environments such as R on a standard laptop. These approaches can be described as model-based solutions for big spatio-temporal datasets. The models ensure that the algorithmic complexity has n floating point operations (flops), where n is the number of spatial locations (per iteration). We compare these methods and provide some insight into their methodological underpinnings.

E0163: **Bayesian regression models for big spatially or longitudinally correlated functional data**
Presenter: **Jeff Morris**, MD Anderson Cancer Center, United States
Co-authors: Lin Zhang, Hojin Yang, Wonyul Lee, Hongxiao Zhu, Veerabhadran Baladandayuthapani

A series of regression modeling strategies that can be used for high-dimensional spatially- or longitudinally correlated functional data will be described. Intrafunctional correlation is handled through basis function modeling, while interfunctional correlation is captured by one of three approaches: (1) parametric or nonparametric random effect functions, (2) separable or non-separable spatial (or temporal) inter-functional processes, or (3) tensor-basis function modeling. Rigorous Bayesian inference is done in such a way that adjusts for any potential multiple testing issues. We will describe these general approaches and illustrate them on a series of complex, high-dimensional, spatially and longitudinally correlated functional data sets coming from strain tensor data from a glaucoma study, bladder cancer genomic maps and event-related potential data from a smoking cessation study. We will also discuss recent work in which we have developed spatiotemporal quantile functional regression approaches that we are applying to model temporal climate change in terms of intraseasonal temperature and precipitation distributions.

E0162: **A general framework for Vecchia approximations of Gaussian processes**
Presenter: **Joseph Guinness**, NC State University, United States
Co-authors: Matthias Katzfuss

Gaussian processes (GPs) are commonly used as models for functions, time series, and spatial fields, but they are, in general, computationally infeasible for large datasets. Focusing on the typical setting of observations from a GP containing a nugget or noise term, we propose a generalization of Vecchia's approximation as a framework for GP approximations. We show that our general Vecchia approach contains many popular existing GP approximations as special cases, allowing a comparison of the different approaches within a unified framework. Representing the models by directed acyclic graphs, we determine the sparsity of the matrices necessary for inference, which leads to new insights regarding the computational properties. Based on these results, we propose a new sparse general Vecchia approximation, which ensures computational feasibility for large datasets but can lead to tremendous improvements in approximation accuracy over Vecchia's original approach. We provide several theoretical results and conduct numerical comparisons. We conclude with guidelines for the use of Vecchia approximations.

EO443 Room CLO B01 BAYESIAN MODEL SELECTION OF GRAPHICAL MODELS Chair: Francesco Stingo

E0410: **Bayesian structure learning in high-dimensional graphical models with application to brain connectivity**
Presenter: **Reza Mohammadi**, Tilburg Univeristy, Netherlands
Co-authors: Helene Massam

The problem of Bayesian structure learning in high-dimensional graphical models is considered, motivated by brain connectivity applications. In graphical models, Bayesian frameworks provide a straightforward tool, explicitly incorporating underlying graph uncertainty. In principle, the Bayesian approaches are based on averaging the posterior distributions of the quantity of interest, weighted by their posterior graph probabilities. However, Bayesian inference has not been used in practice for high-dimensional graphical models, because computing the posterior graph probabilities is hard and the number of possible graph models is very large. We discuss the computational problems related to Bayesian structure learning and we offer several solutions to cope the high-dimensionality problems. We apply our method to high-dimensional fMRI data from brain connectivity studies to show its empirical usefulness. In addition, we have implemented our method in the R package BDgraph which is available online.

E1250: **Bayesian multi-layered Gaussian graphical models**
Presenter: **Min Jin Ha**, UT MD Anderson Cancer Center, United States
Co-authors: Francesco Stingo, Veerabhadran Baladandayuthapani

Simultaneous modeling of data arising from multiple ordered layers provides insight into a holistic picture of the interactive system and the flow of information. Chain graphs have been used to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers that exhibit undirected and directed acyclic relations within and between the layers. We use the multi-layered Gaussian graphical model (mlGGM) to describe a conditional independence structure on a chain graph and propose a Bayesian node-wise selection (BANS) method that coherently accounts for dependencies in the mlGGM. Using variable selection priors for each of the node-wise regressions allows for flexible modeling and the incorporation of edge-specific prior knowledge. Through simulation data generated from various mlGGMs, we demonstrate that our node-wise regression method outperforms other related multivariate regression-based methodologies. We apply BANS to identify integrative networks for key signaling pathways in kidney cancer and dynamic signaling networks using longitudinal protein data from a breast cancer cell line.

E0979: **Bayesian selection of graphical regulatory models**
Presenter: **Silvia Liverani**, Brunel University London, United Kingdom
Co-authors: Jim Smith

A new class of coloured graphical models, called regulatory graphs, is defined. These graphs have their own distinctive formal semantics and can directly represent typical qualitative hypotheses about regulatory processes like those described by various biological mechanisms. They admit an embellishment into classes of probabilistic statistical models and so standard Bayesian methods of model selection can be used to choose promising candidate explanations of regulation. Regulation is modeled by the existence of a deterministic relationship between the longitudinal series of observations labeled by the receiving vertex and the donating one. This class contains longitudinal cluster models as a degenerate graph. Edge colours directly distinguish important features of the mechanism like inhibition and excitation and graphs are often cyclic. With appropriate

distributional assumptions, because the regulatory relationships map onto each other through a group structure, it is possible to define a conditional conjugate analysis. This means that even when the model space is huge it is nevertheless feasible, using a Bayesian MAP search, to a discover regulatory network with a high Bayes Factor score. We also show that, like the class of Bayesian Networks, regulatory graphs also admit a formal but distinctive causal algebra. We illustrate our methods on a microarray experiment measuring the expression of thousands of genes as a longitudinal series.

E1800: **Bayesian graphical regression**
Presenter: **Francesco Stingo**, University of Florence, Italy
Co-authors: Veerabhadran Baladandayuthapani
The purpose is to consider the problem of modeling conditional independence structures in heterogeneous data in the presence of additional subject-level covariates termed Graphical Regression is considered. We propose a novel specification of a conditional (in)dependence function of covariates which allows the structure of a directed graph to vary flexibly with the covariates; imposes sparsity in both edge and covariate selection; produces both subject-specific and predictive graphs; and is computationally tractable. We illustrate our approach in a cancer genomics-based precision medicine paradigm, where-in we explore gene regulatory networks in multiple myeloma taking prognostic clinical factors into account to obtain both population-level and subject-level gene regulatory networks.

EO350 Room MAL B18 CAUSAL INFERENCE IN THEORY AND PRACTICE I Chair: Joris Mooij

E0846: **Structure learning of linear Gaussian structural equation models with weak edges**
Presenter: **Marco Eigenmann**, ETH Zurich, Switzerland
Co-authors: Marloes Maathuis, Preetam Nandy
Structure learning of linear Gaussian structural equation models with weak edges is considered. Since the presence of weak edges can lead to a loss of edge orientations in the true underlying CPDAG, we define a new graphical object that can contain more edge orientations. We show that this object can be recovered from observational data under a type of strong faithfulness assumption. We present a new algorithm for this purpose, called aggregated greedy equivalence search (AGES), that aggregates the solution path of the greedy equivalence search (GES) algorithm for varying values of the penalty parameter. We prove consistency of AGES and demonstrate its performance in a simulation study and on single cell data. The algorithm will be made available in the R-package pcalg.

E1187: **Multi-source causal discovery from real-world experiments with extended JCI**
Presenter: **Tom Claassen**, Radboud University Nijmegen, Netherlands
The focus is on an extension of the recently introduced Joint Causal Inference (JCI) framework for causal discovery from multiple data sets. The framework allows us to jointly learn both the causal structure and targets of interventions from statistical independencies in pooled data. Being able to exploit the information and background knowledge from multiple observational and experimental data sets simultaneously provides for a significant improvement in the accuracy and identifiability of the predicted causal relations, while the systematically pooled data also increases the statistical power of independence tests. Previous implementations of JCI were based on powerful SAT solver approaches, which are very robust and flexible, but unfortunately scale poorly, restricting application to relatively small models. This novel adaptation shows how to extend the JCI framework to standard constraint-based algorithms such as FCI+, which makes it possible to handle both soft and hard (perfect) interventions, as well as larger models up to dozens of variables. We will demonstrate the method on a realistic dynamical simulation of interventions on a gene regulatory network, as well as on several real-world data sets.

E1305: **Learning directed acyclic graphs with hidden variables via latent Gaussian graphical model selection**
Presenter: **Benjamin Frot**, ETH Zurich, Switzerland
Co-authors: Preetam Nandy, Marloes Maathuis
A new method is introduced to estimate the Markov equivalence class of a directed acyclic graph (DAG) in the presence of hidden variables, in settings where the underlying DAG among the observed variables is sparse, and there are a few hidden variables that have a direct effect on many of the observed ones. Building on the so-called low-rank plus sparse framework, we suggest a two-stage approach which first removes unwanted variation using latent Gaussian graphical model selection, and then estimates the Markov equivalence class of the underlying DAG by applying GES. This approach is consistent in certain high-dimensional regimes and performs favourably when compared to the state of the art, both in terms of graphical structure recovery and total causal effect estimation.

E1309: **Nested Markov models**
Presenter: **Ilya Shpitser**, Johns Hopkins University, United States
Co-authors: Thomas Richardson, Robin Evans, James Robins
Directed acyclic graph (DAG) models may be characterized in at least four different ways: via a factorization, the d-separation criterion, the moralization criterion, and the local Markov property. As pointed out previously, marginals of DAG models also imply equality constraints that are not conditional independences. The well-known Verma constraint is an example. We will describe, via a factorization and Markov properties, a statistical model that captures all equality constraints implied by a hidden variable DAG model, including Verma constraints, but not inequality constraints. A previous characterization of these constraint gives an alternative definition of the model. We will also show that the fixing operation used to define this model, an operation that generalizes conditioning, marginalization and applications of the g-formula, gives a particular simple characterization of identifiable causal effects in hidden variable graphical causal models.

EO587 Room MAL B20 STATISTICAL METHODS FOR MULTIPLE RISKS Chair: Holger Rootzen

E0769: **Statistical inference for a relative risk measure**
Presenter: **Yi He**, Monash University, Australia
Co-authors: Yi He, Liang Peng, Yanxi Hou, Jiliang Sheng
For monitoring systemic risk from regulators' point of view, a relative risk measure is proposed, which is sensitive to the market comovement. The asymptotic normality of a nonparametric estimator and its smoothed version is established when the observations are independent. In order to effectively construct an interval without complicated asymptotic variance estimation, a jackknife empirical likelihood inference procedure based on the smoothed nonparametric estimation is provided with a Wilks type of result in case of independent observations. When data follow from AR-GARCH models, the relative risk measure with respect to the errors becomes useful and so we propose a corresponding nonparametric estimator. A simulation study and real-life data analysis show that the proposed relative risk measure is useful in monitoring systemic risk.

E0696: **Regularly varying time series and max-stable processes**
Presenter: **Anja Janssen**, KTH Royal Institue of Technology, Sweden
Some connections between stationary multivariate regularly varying time series and the class of max-stable processes in discrete time will be discussed. We focus on the extremal aspects of both processes which can be expressed in terms of a limiting process that is known as the (spectral)

tail process, which describes the conditional behavior of the process before and after an extremal event at a given time. We explore how the stationarity of the processes is reflected in the extremal behavior and how it is for example possible to construct an underlying stationary max-stable process for a given spectral tail process.

E0760: **On the random number of multivariate risks**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Enkelejd Hashorva, Stefano Rizzelli
Modeling of joint extreme-values is of interest in many real applications. Consider for simplicity a bivariate random vector (X_1, X_2), with joint distribution function F, representing two quantities of interest and let $(X_{1,1}, X_{1,2})$, ..., $(X_{n,1}, X_{n,2})$ be iid copies of it. Let N be a non-negative random variable representing the total number of events that have occurred during the period under investigation. Recent contributions investigate the extremal dependence underlying the distribution function, denoted by H, of the compound random variables $\left(\max_{1\leq i\leq N} X_{i,1}, \max_{1\leq i\leq N} X_{i,2}\right)$. Under appropriate conditions on N, it has been shown that the extremal properties of the distribution H are the same as those of the distribution F. An open question we are interested in is the reverse problem. Specifically, we address the following question: what are the extremal properties of F given that those of H are known? We answer this question by exploiting the multivariate extreme-value theory. We provide the conditions on N under which extremal properties of H are different from those of F. Starting from an estimator of the extremal dependence of H we derive an estimator for the extremal dependence of F. For the latter estimator we derive the asymptotic properties and for finite samples its performance is illustrated via a simulation study.

E1349: **The value of a liability cash flow in discrete time subject to capital requirements**
Presenter: **Filip Lindskog**, Stockholm University, Sweden
The aim is to define the market-consistent value of a liability cash flow in discrete time subject to repeated capital requirements, and explore its properties. Our multi-period market-consistent valuation approach is based on defining a criterion for selecting a static replicating portfolio and defining the value of the residual liability, whose cash flow is the difference between the original liability cash flow and that of the replicating portfolio. The value of the residual cash flow is obtained as a solution to a backward recursion that is implied by the procedure for financing the repeated capital requirements, and no-arbitrage arguments. We show that the liability value resulting from no-arbitrage pricing of the dividends to capital providers may be expressed as a multi-period cost-of- capital valuation. Explicit valuation formulas are obtained under Gaussian model assumptions.

EO619 Room MAL B30 ADVANCES IN STATISTICAL METHODS IN SURVIVAL ANALYSIS AND MISSING DATA Chair: Sebastien Haneuse

E0790: **Regression analyses of survival data subject to biased sampling**
Presenter: **Yu Shen**, UT MD Anderson Cancer Center, United States
Methodologic development in semiparametric modeling of length-biased data has made considerable progress in recent years in many different directions. We will give an overview on recent semiparametric modeling for right-censored survival data under length-biased sampling. The methods will be reviewed for commonly used proportional hazards model, and AFT model for time-to-event outcomes, and restricted mean survival times. The estimation methods cover both estimating equation approaches and full likelihood methods using EM algorithm. Some related software for the implementation of such methods will be illustrated.

E1168: **Exact inference on the restricted mean survival time**
Presenter: **Lu Tian**, Stanford University, United States
In a randomized clinical trial with the time to event as the primary endpoint, one often evaluates the treatment effect by comparing the survival distributions from two groups. This can be achieved by for example estimating the hazard ratio under the popular proportional hazards (PH) model. However, when the hazard rate is very low, e.g., in safety studies, there may be too few observed events to warrant the valid asymptotic inferences based on the PH model. The exact inference including hypothesis testing and constructing 95% confidence interval for the treatment effect is desired. We have developed an exact inference procedure for estimating the treatment effect based on the difference in restricted mean survival time between two arms, which is more appealing than hazard ratio in many applications. The proposed procedure is valid regardless of the number of events. We have also performed a simulation study to examine the finite sample performance of the proposed method.

E1339: **Survival analysis with presence of informative censoring via nonparametric multiple imputation**
Presenter: **Chengcheng Hu**, University of Arizona, United States
Co-authors: Jeremy Taylor, Chiu-Hsieh Hsu
A nonparametric multiple imputation approach is developed to recover information for censored observations while analyzing survival data with presence of informative censoring. A working shared frailty model is proposed to estimate the magnitude of informative censoring through estimating Kendall's tau, which is only used to determine the size of imputing risk set for each censored subject. Specifically, a larger tau indicates a smaller size of the imputing risk set. It has been shown that the posterior mean of frailty is a monotonic function of the observed time under the assumed models. Therefore, the observed times for subjects at risk are used to determine the imputing risk set for each censored subject. Simulation studies have shown that the nonparametric multiple imputation approach produces survival estimates close to the targeted values and coverage rates of the confidence intervals close to the nominal level, even in situations with a high degree of informative censoring.

E1307: **Inference for high-dimensional models**
Presenter: **Yi Li**, University of Michigan, United States
Inference for high-dimensional models is challenging as penalization and selection are often involved. A novel way of simultaneous estimation and inference for high-dimensional linear models is proposed. By smoothing over partial regression estimates based on some variable selection scheme, we reduce the problem to a low-dimensional case such as fitting least squares. The procedure, termed as Selection-assisted Partial Regression and Smoothing (SPARES), utilizes data-splitting with bootstrap, variable selection, and partial regression. We show that the SPARES estimator is asymptotically unbiased and normal, and derive its variance via a non-parametric delta method. The utility of the procedure is evaluated under various simulation scenarios and via comparisons with some competing methods, the de-biased LASSO estimators. We apply the method to analyze two genomic datasets.

EO138 Room MAL B33 RECENT ADVANCES IN FDA, HIGH DIMENSIONAL AND SPATIAL STATISTICS Chair: Graciela Boente

E0314: **A measure of directional outlyingness with applications to image data and video**
Presenter: **Jakob Raymaekers**, KU Leuven, Belgium
Co-authors: Peter Rousseeuw, Mia Hubert
Images and video can be considered as functional data with a bivariate domain, where the data per grid point can be univariate (e.g. grayscale values) or multivariate (e.g. red, green, and blue intensities). This often yields large datasets, in which outliers may occur that can distort the analysis. At each grid point we propose to compute a fast measure of outlyingness which accounts for skewness in the data. It can be used for univariate data and, by means of projection pursuit, for multivariate data. The influence function of this outlyingness measure is computed as well as its implosion and explosion bias. We also construct a cutoff value for the outlyingness. Heatmaps of the outlyingness indicate the regions in which an image deviates most from the majority of images. To illustrate the performance of the method it is applied to real multivariate functional data. One example consists of MRI images which are augmented with their gradients. We also show an example of video surveillance data, where we compare the exact method with faster approximations.

E0829: **Screening for ultrahigh-dimensional regression with cellwise outliers**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Yixin Wang
In ultrahigh-dimensional data it can easily happen that most or all of the observations are contaminated in some of their cells. To robustly estimate a best approximating subspace in this setting, we consider componentwise LTS-estimators. We propose an efficient algorithm to calculate these estimators by using estimating equations and deterministic starting values. We apply these methods in a robust variable selection procedure for ultra-high dimensional regression analysis. In particular, we propose a robust version of Factor Profiled Sure Independence Screening. By assuming that the predictors can be represented by a few latent factors, this method can handle correlation among the candidate predictors. We use robust componentwise LTS-estimators to estimate the factors. Then, a robust regression method is applied on the profiled variables to screen for the most important predictors.

E0763: **Penalized M-estimators in logistic regression**
Presenter: **Ana Maria Bianco**, Universidad de Buenos Aires, Argentina
Co-authors: Graciela Boente, Gonzalo Chebi
The problem of variable selection in logistic regression is crucial when the number of variables is high. We introduce a family of penalized M-type estimators for logistic regression that are stable against atypical data. Theoretical results regarding oracle properties are studied. A numerical study that illustrates the finite sample behaviour of the proposal is presented.

E0698: **Prediction based on multivariate spatial data: A sufficient dimension reduction approach**
Presenter: **Pamela Llop**, Facultad de Ingenieria Quimica, UNL-CONICET, Argentina
Co-authors: Liliana Forzani, Maria Antonella Gieco, Anne Francoise Yao
One of the principal objectives in spatial statistics is to reconstruct certain phenomenon of interest, using data measured over its region domain. The best known methods to perform such reconstruction are based mainly in predictions that use not only the variable of interest (response) but also extra variables (predictors) which are added to the model in order to get better results. When the amount of extra variables is large, it may be of interest to reduce it in order to simplify the analysis but without losing information about the phenomenon to be studied. The sufficient dimension reduction techniques (SDR) consists in reducing the high dimensional space of predictors combining them in a new set of variables that lives in an lower dimensional space without losing information about the response. We apply SDR techniques in order to perform prediction based on multivariate spatially correlated data. The good performance of our method is shown via some simulation studies and application to real examples.

EO178 Room MAL B34 ADVANCES IN FUNCTIONAL AND HIGH-DIMENSIONAL DATA ANALYSIS Chair: Jeng-Min Chiou

E0575: **Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippets**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Longitudinal data are often plagued with sparsity of time points where measurements are available. The functional data analysis perspective has been shown to provide an effective and flexible approach to address this problem for the case where measurements are sparse but their times are randomly distributed over an interval. We focus here on a different scenario where available data can be characterized as snippets, which are very short stretches of longitudinal measurements. For each subject the stretch of available data is much shorter than the time frame of interest, a common occurrence in accelerated longitudinal studies. An added challenge is introduced if a time proxy that is basic for usual longitudinal modeling is not available, as encountered in situations where time of disease onset is unknown and chronological age does not provide a meaningful time reference for longitudinal modeling. To address these challenges, we introduce conditional quantile trajectories for monotonic processes as solutions of a dynamic system. Conditional quantile trajectories emerge as useful descriptors of processes that quantify deterioration over time.

E0194: **A functional dependence measure for large curve time series with an application to autoregressions**
Presenter: **Xinghao Qiao**, London School of Economics, United Kingdom
Co-authors: Shaojun Guo
Modelling a large bundle of curves arises in a broad spectrum of real applications. However, many studies in functional data analysis literature focus primarily on the critical assumption of independent and identically distributed samples of a fixed number of curves. We introduce a measure of functional dependence for stationary functional processes that provides insights into the effect of cross-dependence among high dimensional curve time series. Based on our proposed functional dependence measure, we establish some useful concentration bounds for the relevant estimated terms when each component of the vector of curve time series is represented through its Karhunen-Loève expansion. As an example to illustrate, we propose vector functional autoregressive models, which characterize the dynamic dependence across high dimensional curve time series, and develop a regularization approach to estimate autoregressive coefficient functions. We then apply our developed concentration bound results to derive the non-asymptotic upper bounds for the estimation errors of the regularized estimates. We also show that the proposed method significantly outperforms its potential competitors through a series of simulated experiments and one real world data example. Finally, we discuss the application of our proposed functional dependence measure on some possible topics, e.g. causality and factor models.

E0431: **Weak separability and L-separability for multi-way functional data**
Presenter: **Kehui Chen**, University of Pittsburgh, United States
Multi-way functional data refers to functional data with double or multiple indices, such as brain-imaging data with spatial and temporal indices. To achieve efficient dimension reduction, one usually adopts the strong separability assumption that the covariance is a product of a spatial covariance and a temporal covariance. This assumption is quite restrictive and is often violated in real applications. We will introduce weak separability and L-separability, where the covariance can be approximated by a weighted sum of strong separable components. We will present the formal test procedure for the weak separability assumption and the algorithm for L-separable decomposition.

E1169: **Linear classification for functional data with direct estimation**
Presenter: **Juhyun Park**, Lancaster University, United Kingdom
Co-authors: Jeongyoun Ahn, Yongho Jeon
Functional data are inherently infinite-dimensional and, thus, dimension reduction is crucial to solve many inverse problems arising in statistical analysis. Functional PCA has been popular as a key technique to find an efficient finite dimensional representation. Many regression and clustering solutions are based on that, as essentially the inverse of the covariance operator is well defined. However, it is known that functional classification can achieve a perfect classification, if high dimensionality is well exploited. Hence, for the purpose of classification, it is not necessarily advantageous to have a well-defined finite-dimensional representation. An alternative method such as partial least squares method may find a better representation to exploit the dimensionality. Nevertheless, selecting the truncation order to define a finite representation in general is not a trivial issue. Based on these observations, we seek an alternative approach to classification with a direct estimation method. We consider the problem in linear methods and formulate it as a regularization problem with appropriate penalty. An added advantage of using penalty formulation is the possibility of incorporating some structural constraints in functional data such as sparsity or smoothness as we desire. We study the performance of the new method and develop an efficient algorithm.

EO605 Room Bloomsbury RECENT DEVELOPMENT IN GENETIC EPIDEMIOLOGY Chair: Malka Gorfine

E0495: **Recommendation of when to treat: From binary to time-to-intervention decision**
Presenter: **Li Hsu**, Fred Hutchinson Cancer Research Center, United States
Precision medicine has the potential to improve the practice of disease prevention and treatment. For many complex diseases, lifestyle, environmental, owing to high throughput omics technologies, many genetic risk factors have already been identified. This has raised the expectation that the risk prediction models built upon these risk factors can substantially improve the prediction accuracy. It is thus important to understand how the model can be used in clinical practice. It is common to use the model to make a binary decision, e.g., whether or not a test should be offered given the subjects risk profile. Many measures have been proposed to evaluate the usefulness of a model with such a binary decision. However, sometimes it is also of interest to know when to treat. We will present a novel concept, recommended time to start intervention based on the subjects risk profiles and the time-dependent risk prediction model. We will also present time-dependent measures for assessing the usefulness of the model with the when-to- treat decision. This will add to the tools that patients, providers and policy makers can use to make individualized decisions, which ultimately improve the patients health without unnecessary treatments or tests.

E0455: **Quantification of multiple tumor clones using gene array data or sequencing data**
Presenter: **Charles Kooperberg**, Fred Hutchinson Cancer Research Center, United States
It is not uncommon that tumor samples are in fact mixtures of several "clones" and possible also some normal cells. We describe a maximum likelihood based algorithm for determining what percentage of the tumor sample particular clones are from, as well as the copy number alterations of each of these clones, using nowadays quite economical gene array data or sequencing data. We illustrate our results on a cell-line data set of Barrett's esophagus, a pre-curser of esophageal cancer.

E1444: **On the utility of a comprehensive family history of colorectal cancer improve risk prediction**
Presenter: **Yingye Zheng**, Fred Hutchinson Cancer Research Cener, United States
Family history of colorectal cancer (CRC) is a strong and well-established risk factor for CRC. To date, however, family history (FH) of the disease is generally only broadly categorized (usually as present or absent) in most risk prediction models. We investigated the utility of a CRC risk model that incorporates a comprehensive family history of CRC as well as information on known genetic and environmental risk factors and personal characteristics. We used data from the Colon Cancer Family Registry (CCFR). A familial risk profile (FRP) score, a probability index of absolute risk for lifetime CRC was estimated based on family structure, age of onset for affected relatives and the polygenic effect of MLH1, MSH2, MSH6, PMS2 and MUTYH using *modified segregation analysis*, an adapted previous approach used as a measure of family history in the risk model. The primary endpoint was CRC diagnosis within 5-year after baseline. We used calibration plots to compare the predicted 5-year absolute risks with the observed cumulative incidence rates. Receiver Operating Characteristic (ROC) curve analyses were conducted to assess the discriminatory capacity for separating subjects with and without a CRC diagnosis within 5 years, accounting for censoring and competing risk.

E1796: **Illness-death and marginalized frailty models with application to incident and prevalent event data**
Presenter: **Malka Gorfine**, Tel Aviv University, Israel
A cross-sectional sampling design is considered which is limited to a certain age group (e.g. 40-70), with a follow-up period until death or censoring, whichever comes first. The various sources of possible bias, in estimating the covariates effect on survival, will be discussed. Unbiased estimators will be presented, based on illness-death frailty-based models. The models and methods will be applied using Biobank data with genetic variants as covariates.

EO043 Room Chancellor's Hall THEORETICAL FOUNDATIONS OF BIG DATA Chair: Will Wei Sun

E0557: **Unsupervised ensemble learning**
Presenter: **Boaz Nadler**, Weizmann Institute of Science, Israel
In various contemporary applications, one is given the advice or predictions of several classifiers of unknown reliability, over multiple questions or queries. This scenario is different from standard supervised learning where classifier accuracy can be assessed from available labeled training or validation data, and raises several questions: given only the predictions of several classifiers of unknown accuracies, over a large set of unlabeled test data, is it possible to a) reliably rank them, and b) construct a meta-classifier more accurate than any individual classifier in the ensemble? we show that under various independence assumptions between classifier errors, this high dimensional data hides simple low dimensional structures. Exploiting these, we present simple spectral methods to address the above questions, and derive new unsupervised spectral meta-learners. We prove these methods are asymptotically consistent when the model assumptions hold, and present their empirical success on a variety of unsupervised learning problems.

E0583: **Renewable maximum likelihood estimation in generalized linear models for streaming data**
Presenter: **Peter Song**, University of Michigan, United States
Co-authors: Lan Luo
A new incremental learning algorithm is presented to analyze streaming data using the generalized linear models. The proposed method is developed within a new framework of renewable estimation, in which the maximum likelihood estimation can be renewed with current data and summary statistics of historic data, but with no use of any historic data themselves. In the implementation, we design a new data flow, called the rho architecture to accommodate the data storage of current and historic data, as well as to communicate with the computing layer of the system in order to facilitate sequential learning. We prove both estimation consistency and asymptotic normality of the renewable MLE, and propose some sequential inferences for model parameters. We illustrate our methods by various numerical examples from both simulation experiments and real-world analysis.

E0787: **Learning large-scale Bayesian networks**
Presenter: **Qing Zhou**, UCLA, United States
Co-authors: Bryon Aragam, Arash Amini, Jiaying Gu
Learning graphical models from data is an important problem with wide applications, ranging from genomics to the social sciences. Nowadays datasets typically have upwards of thousands, sometimes tens or hundreds of thousands, of variables and far fewer samples. To meet this challenge, we develop theory and algorithms for learning the structure of large Bayesian networks, represented by directed acyclic graphs (DAGs). Our theoretical results establish support recovery guarantees and deviation bounds for a family of penalized least-squares estimators under concave regularization, including many popular regularizers, such as the MCP, SCAD, ℓ_0 and ℓ_1. The proof relies on interpreting a DAG as a recursive linear structural equation model, which reduces the estimation problem to a series of neighborhood regressions. We apply these results to study the statistical properties of score-based DAG estimators, learning causal DAGs, and inferring conditional independence relations via graphical models. Our algorithms are implemented in a new open-source R package, sparsebn, available on CRAN. This package focuses on the unique setting of learning large networks from high-dimensional data, possibly with interventions, and places a premium on scalability.

E1831: **Consistent change-point detection and parameter estimation in high-dimensional piecewise-stationary VAR models**
Presenter: **Ali Shojaie**, University of Washington, United States
Co-authors: Abolfazl Safikhani
Assuming stationarity is unrealistic in many time series applications. A more realistic alternative is to allow for piecewise stationarity, where the model is allowed to change at given time points. We consider the problem of detecting the change points and estimation of model parameters in a high-dimensional piecewise vector autoregressive model (VAR). To this end, we propose a two-stage estimation strategy for consistent estimation of both the change points, as well as the parameters of the VAR process. In the first step, we reformulate the change point detection problem as a high-dimensional variable selection one, and propose a penalized least square estimator using a total variation penalty. We show that the proposed penalized estimation method over-estimates the number of change points. We thus propose a backward selection criterion in conjunction with the penalized least square estimator to tackle this issue. We prove that the proposed two-stage procedure consistently detects the number of change points and their locations. We also show that the procedure results in consistent estimation of VAR parameters. A block coordinate descent algorithm is developed for efficient computation of model parameters. The performance of the method is illustrated using several simulation scenarios, and real data examples.

EO132 Room Court DEPENDENCE MODELS AND COPULAS FOR CLIMATE AND ENVIRONMENT Chair: Fabrizio Durante

E0178: **Modeling spatial processes with unknown extremal dependence class**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Jenny Wadsworth
Many environmental processes exhibit weakening spatial dependence as events become more extreme. Well-known limiting models, such as max-stable or generalized Pareto processes, cannot capture this, which can lead to a preference for models that exhibit a property known as asymptotic independence. However, weakening dependence does not automatically imply asymptotic independence, and whether the process is truly asymptotically (in)dependent is usually far from clear. The distinction is key as it can have a large impact upon extrapolation, i.e., the estimated probabilities of events more extreme than those observed. We present a single spatial model that is able to capture both dependence classes in a parsimonious manner, and with a smooth transition between the two cases. The model covers a wide range of possibilities from asymptotic independence through to complete dependence, and permits weakening dependence of extremes even under asymptotic dependence. Censored likelihood-based inference for the implied copula is feasible in moderate dimensions due to closed-form margins. The model is applied to oceanographic datasets with ambiguous true limiting dependence structure. We also show how a bivariate restriction of this copula model may be used to dynamically capture the extremal dependence structure of financial data over time.

E0273: **Nonparametric estimation of multivariate quantiles in small sample sizes**
Presenter: **Maximilian Coblenz**, Karlsruhe Institute of Technology, Germany
Co-authors: Oliver Grothe, Rainer Dyckerhoff
In many applications, quantiles provide important insights in the statistical problems considered. We focus on the estimation of multivariate quantiles based on copulas. We provide a nonparametric estimation procedure for a specific notion of multivariate quantiles which is in particular useful for applications in hydrology. These quantiles are based on level sets of copulas and admit the usual probabilistic interpretation that a p-quantile comprises a probability mass p. We design a smoothed bootstrap procedure to cope with the typically small sample sizes expected in hydrology. We show that, in particular for extreme events with return periods of 30 years and more, the proposed procedure highly improves precision and accuracy of the estimation results compared to base line approaches.

E0258: **An application of copulae in the analysis of sand transport phenomena**
Presenter: **Franco Pellerey**, Politecnico di Torino, Italy
Co-authors: Lorenzo Raffaele, Luca Bruno, Luigi Preziosi
The reliable prediction of the consequences of aeolian events related to sand transport phenomena is a key element for human activities in many arid regions of the word, and threshold shear velocity generating sand lifting is a key component of such a prediction. To this aim, a list of deterministic semi empirical models have been proposed in the specialized literature. Unfortunately, such models suffer from the effects of uncertainties of different origins, such as those related to physical phenomena or measurement procedures, and are often fitted to a small amount of data. Probabilistic models have been recently proposed as well, but they can be hardly applied since they require the knowledge of the distributions of a large number of random factors. The aim is to describe a purely statistical approach able to provide estimates of fluid threshold shear velocity for sand saltation based on a data set derived from published studies. In this approach, the relationship existing between threshold shear velocity and grain diameter of the sand is described through a suitable copula. Estimates of conditional probability distributions are shown and critically compared to those proposed in previous studies.

E0980: **Discrete copulas for weather forecasting: Theoretical and practical aspects**
Presenter: **Elisa Perrone**, Massachusetts Institute of Technology, United States
Co-authors: Liam Solus, Caroline Uhler
Discrete copulas are fascinating geometric objects of great importance for empirical modeling in the applied sciences. We analyze mathematical and practical features of discrete copulas, i.e., restrictions of copula functions to non-square grid domains. In particular, we (1) highlight fundamental connections between copulas and discrete geometry, and (2) discuss their application to weather forecasting problems. First, we present a geometric approach to describe families of discrete copulas through the properties of their associated polytopes. Specifically, we inquire into the geometric attributes of these polytopes such as bounding hyperplanes, vertices, and volume. Then, we discuss the role of discrete copulas in empirical postprocessing methods of numerical weather ensemble forecasts by presenting a case study on weather data of the Austrian region.

EO332 Room G11 ADVANCED COMPUTATIONAL METHODS FOR STATISTICAL MODELING Chair: Tsung-I Lin

E0184: **Heavy-tailed longitudinal regression models for censored data: A likelihood based perspective**
Presenter: **Larissa Matos**, Campinas State University - UNICAMP, Brazil
Co-authors: Tsung-I Lin, Mauricio Castro, Victor Hugo Lachos Davila
HIV RNA viral load measures are often subjected to some upper and lower detection limits depending on the quantification assays. Hence, the responses are either left or right censored. Moreover, it is quite common to observe viral load measurements collected irregularly over time. A complication arises when these continuous repeated measures have a heavy-tailed behaviour. For such data structures, we propose a robust nonlinear censored regression model based on the scale mixtures of normal (SMN) distributions. To take into account the autocorrelation existing among irregularly observed measures, a damped exponential correlation structure is considered. A stochastic approximation of the EM (SAEM) algorithm is developed to obtain the maximum likelihood estimates of the model parameters. The main advantage of this new procedure is that it allows us to estimate the parameters of interest and evaluate the log-likelihood function in an easy and fast way. Furthermore, the standard errors of the fixed effects and predictions of unobservable values of the response can be obtained as a by-product. The practical utility of the proposed method is exemplified using both simulated and real data.

E0550: **On moments of truncated multivariate Student-t distribution: A recurrence approach**
Presenter: **Christian Galarza**, State University of Campinas, Brazil
Co-authors: Victor Hugo Lachos Davila, Tsung-I Lin, Wan-Lun Wang
Recurrence relations for integrals that involve the density of multivariate Student-t distributions are developed. These recursions allow fast computation of the moments of folded and truncated multivariate normal and Student-t distributions. Besides being numerically efficient, the proposed recursions also allow us to obtain explicit expressions of low order moments of folded and truncated multivariate Student-t distributions. The newly methods are implemented in the new R package MoMtt.

E0804: **Censored time series analysis for responses on the unit interval: An application to acid rain modeling**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Fernanda Schumacher, Victor Hugo Lachos Davila, Guillermo Ferreira
An autoregressive model is proposed for time series in which the variable of interest lie on the unit interval being subjected to certain threshold values below or above which the measurements are not quantifiable. The model includes the independent beta regression as a special case. We provide a full Bayesian approach for the estimation of the model parameters using standard Markov Chain Monte Carlo (MCMC) methods, simulating samples for the joint posterior distribution. We discuss the construction of the proposed model and compare it with alternative models using simulated and real data sets.

E0780: **A pattern-clustering method for longitudinal data on heroin users receiving methadone**
Presenter: **Chien-Ju Lin**, MRC Biostatistics Unit, United Kingdom
Methadone is used as a substitute of heroin and there may be certain groups of users according to methadone dosage. We analyze data for 314 participants over 180 days. The data consist of seven categories in which six categories have an ordinal scale for representing dosages and one for missing dosages. We develop a clustering method involving the so-called p-dissimilarity and an ordering algorithm. The p-dissimilarity is used to measure dissimilarity between the 180-day time series of the participants. It accommodates ordinal and categorical scales by using a parameter p as a switch between data being treated as categorical and ordinal. We use heatplots to evaluate the quality of clustering. A heatplot consists of horizontal lines representing data for participants by colour. The interpretability depends on the location of the participants along the vertical-axis. We propose an algorithm using a projection vector to locate participants. The heatplot can then be used for information visualisation. It displays clustering structures, relationships between participants and clusters, and the density of clusters. Despite the fact that no significant clustering structure is observed, the sequence of categories for clusters are useful for clinicians to prescribe an appropriate dosage for improving efficiency of methadone maintenance therapy.

EO591 Room G3 COMPUTATIONAL STATISTICS IN DISTRIBUTION THEORY Chair: Filipe Marques

E0278: **The simultaneous test of equality and circularity of several covariance matrices**
Presenter: **Filipe Marques**, NOVA.ID.FCT, Portugal
Co-authors: Carlos Coelho
The simultaneously test of equality and circularity of several covariance matrices is studied, and accurate near-exact approximations are developed for the distribution of the associated likelihood ratio test statistic. The equality of covariance matrices test is widely used in many statistical procedures and, recently, matrices with a circular pattern, which are particular cases of the well known Toeplitz matrices, have been considered in many practical applications. Therefore a statistical procedure which may allow to test simultaneously both the equality and circularity of a given number of covariance matrices may be a powerful and useful practical tool. Based on a decomposition of the null hypothesis in study, which induces a convenient factorization on the likelihood ratio test statistic, the authors show how it is possible to derive precise and easy to implement near-exact approximations for the distribution of the likelihood ratio test statistic. Numerical studies are carried out to assess the quality of these approximations and to study the properties of this test.

E0449: **A more general framework for the skew normal distributions**
Presenter: **Andriette Bekker**, University of Pretoria, South Africa
Co-authors: BW Rowland, JT Ferreira, Mohammad Arashi
The normal distribution is popular in many statistical contexts. However it is rather restrictive to apply to real world applications due to its symmetry and tail behaviour. In order to alleviate the aforementioned issues, a generalised normal distribution that exhibits flexibility in its tail behaviour is proposed as candidate to apply existing skewing methodology to. Methods to approximate the characteristics of this skew generalised-normal type I distribution (SGNI) and a corresponding stochastic representation are presented. The existing skewing methodology is extended to the elliptical class with the generalised normal distribution as the symmetric base. Furthermore, this SGNI distribution, along with other competing distributions (normal and skew- normal), is used in a distribution fitting context. In the landscape of skew distributions, this new SGNI distribution outperforms existing competing distributions, specifically in approximating particular binomial distributions when p not equal to 0.5.

E1304: **Testing the adequacy of semiparametric transformation models**
Presenter: **James Allison**, Northwest University, South Africa
Co-authors: Marie Huskova, Simos Meintanis
A semiparametric model is considered whereby the response variable following a transformation can be expressed by means of a regression model. In this model the form of the transformation is specified analytically (up to an unknown transformation parameter), while the regression function is completely unknown. We develop testing procedures for the null hypothesis that this semiparametric model adequately describes the data at hand. In doing so, the test statistic is formulated on the basis of Fourier-type conditional expectations, an idea first put forward by Bierens. The

asymptotic distribution of the test statistic is obtained under the null as well as under alternative hypotheses. Since the limit null distribution is nonstandard, a bootstrap version is utilized in order to actually carry–out the test procedure. Monte Carlo results are included that illustrate the finite-sample properties of the new method.

E0852: **Development and computational implementation of approximations for the product of independent random variables**
Presenter: **Theodor Loots**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Filipe Marques
Various general procedures exist for deriving representations of the exact distribution of the product of independent random variables. Such products are related to the distribution of many LRT statistics used in testing, for example, the equality of covariance matrices (equal samples size case), sphericity, independence of several groups of variables, compound symmetry and circularity. These general results are used to show that these can be used to address the product of particular choices of independent random variables, with the main contribution being the development of an R package showcasing its implementation.

EO520 Room G4 RECENT ADVANCES IN RESEARCH SYNTHESIS METHODS Chair: Yinghui Wei

E1827: **Hierarchical models for combining N-of-1 trials**
Presenter: **Christopher Schmid**, Brown University, United States
N-of-1 trials are single-patient multiple-crossover studies for determining the relative effectiveness of treatments for an individual participant. Although N-of-1 trials are designed to estimate treatment efficacy in single patients, N-of-1 trials assessing the same scientific questions may be combined together and analyzed with a multilevel model so that in aggregate they increase the information about each patient by borrowing of strength. When the treatments compared differ between trials, it may be possible to construct a network of treatments and use network meta-analysis methods to make comparisons among the treatments both at the individual and population levels. We will compare results from network analytical models with those from analyses of individual data alone. We will use data from a series of N-of-1 trials in an ongoing study assessing different treatments for chronic pain to demonstrate the application of hierarchical models for N-of-1 data.

E1828: **A model for meta-analysis of correlated binary outcomes: The case of split-body interventions**
Presenter: **Orestis Efthimiou**, University of Bern, Switzerland
Co-authors: Dimitris Mavridis, Adriani Nikolakopoulou, Gerta Rucker, Sven Trelle, Matthias Egger, Georgia Salanti
In several areas of clinical research it is common for trials to assign different sites of the participants bodies to different interventions. For example, a randomized controlled trial (RCT) comparing surgical techniques for correcting myopia may randomize each eye of a participant to a different operation. Under such bilateral (split-body) interventions, the observations from each participant are correlated. It is challenging to account for these correlations at the meta-analysis level, especially when the outcome is rare. Here we present a meta-analysis model based on the bivariate binomial distribution. Our model can synthesize studies on patients who received one intervention at one body site, patients who received two interventions at different sites, or a mixture of these two groups. The model can analyze studies with zero events in one or both treatment arms and can handle the case of incomplete data reporting. We use simulations to assess the performance of our model and to compare it with the bivariate beta-binomial model. In the case of bilateral interventions our model performed well and outperformed the bivariate beta-binomial model in all scenarios explored. We illustrate our methods using two previously published meta-analyses from the fields of orthopaedics and ophthalmology. We conclude that our model constitutes a useful new tool for the meta-analysis of binary outcomes in the presence of split-body interventions.

E1833: **Towards automated biomedical evidence synthesis: RobitReviewer**
Presenter: **Iain Marshall**, Kingś College London, United Kingdom
Co-authors: Byron Wallace
RobotReviewer is presented, an open-source web-based system that uses machine learning and natural language processing (NLP) to semi-automate biomedical evidence synthesis, to aid the practice of Evidence-Based Medicine. RobotReviewer processes full-text journal articles (PDFs) describing randomized controlled trials (RCTs). It appraises the reliability of RCTs and extracts text describing key trial characteristics (e.g., descriptions of the population) using novel statistical NLP methods. RobotReviewer then automatically generates a report synthesising this information. Our goal is for RobotReviewer to automatically extract and synthesise the full-range of structured data needed to inform evidence-based practice.

E1824: **Multivariate network meta-analysis made simple**
Presenter: **Yong Chen**, Univ. of Pennsylvania, United States
The growing number of treatment options for many conditions has generated an increasing need for scientifically rigorous comparisons of multiple treatments to inform healthcare decision making. There is a steep increase in the publication of network meta-analysis (NMA) in the past decade. NMA expands the scope of conventional pairwise meta-analysis by allowing comparisons of larger sets of treatment options. However, the existing methods development have been focusing on univariate outcome NMA. In practice, efficacy, safety and patient-centered outcomes are crucial for clinical decision making and must be considered simultaneously. Furthermore, different patients can have different preferences (or utility function) on the tradeoff among these multivariate outcomes. We propose a modeling and inference strategy for multiple outcomes network meta-analysis. The proposed method avoids modeling the complex correlation structure among multivariate outcomes, whilst providing the valid inference on personalized utility functions. We also provide a subject-specific treatment ranking based on the surface under the cumulative ranking curve. The proposed method is validated through simulation studies and illustrated by an example of treatment comparisons for bipolar disorder.

EO662 Room G5 EPIGENETIC DATA Chair: Qi Zhang

E1028: **Integrative Bayesian models for precision oncology**
Presenter: **Veerabhadran Baladandayuthapani**, UT MD Anderson Cancer Center, United States
Co-authors: Jeff Morris, Francesco Stingo, Kim-Anh Do
Modern biomedicine has generated unprecedented amounts of data. A combination of clinical, environmental and public health information, proliferation of associated genomic information, and increasingly complex digital information have created unique challenges in assimilating, organizing, analyzing and interpreting such structured as well as unstructured data. Each of these distinct data types provides a different, partly independent and complementary, high-resolution view of various biological processes. Modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. pathway/regulatory mechanisms, serial and spatial correlations etc.). Integrative analyses of these multi-domain data combined with patients clinical outcomes can help us understand the complex biological processes that characterize a disease, as well as how these processes relate to the eventual progression and development of a disease. The aim is to cover statistical and computational frameworks that acknowledge and exploit these inherent complex structural relationships for both biomarker discovery and clinical prediction to aid translational medicine. The approaches will be illustrated using several case examples in oncology.

E0193: **Statistical inference on epi-allelic patterns and profiling of DNA methylation from WGBS data**
Presenter: **Carsten Wiuf**, University of Copenhagen, Denmark

The study of epigenetic heterogeneity at the level of individual cells and in whole populations is the key to understanding cellular differentiation, organismal development, and the evolution of cancer. We have developed a statistical method (epiG) to infer and differentiate between various epi-allelic haplotypes, annotated with CpG methylation status and DNA polymorphisms, from WGBS data, and nucleosome occupancy from NOMe-seq data. It is a likelihood-based method that clusters WGBS reads into epi-allelic haplotypes based on sequence similarity, while taking into account experimental errors and biological noise. It outputs the dominating epi-allelic haplotypes of a genomic region of interest, annotated with methylation profiles. The capabilities of the method are demonstrated by inferring allele-specific methylation and nucleosome occupancy in cell lines, and colon and tumor samples, and by benchmarking the method against independent experimental data.

E1058: **Nonparametric empirical Bayes mixture models in Hi-C peak calling and Allelic Imbalance detection from ChIP-seq**
Presenter: **Qi Zhang**, University of Nebraska Lincoln, United States

Over-dispersion is a common phenomenon observed in NGS data. If not properly modeled, it may severely influence the accuracy of statistical inference, e.g., peak calling, detecting differentially expressed genes or allelic imbalance. Commonly used parametric mixture models fail to capture such overdispersion due to the parametric constraint. We propose an non-parametric empirical Bayes framework for modeling NGS data, and we apply it to the problems of detecting allelic imbalance from ChIP-Seq data and calling peaks from HiC data.

E1535: **Statistical methods for profiling 3-dimensional chromatin interactions from repetitive regions of genomes**
Presenter: **Ye Zheng**, University of Wisconsin Madison, United States
Co-authors: Ferhat Ay, Sunduz Keles

Recently developed chromatin conformation capture-based assays enabled the study of 3-dimensional chromosome architecture in a high throughput fashion. Hi-C, particularly, elucidated genome-wide long-range interactions among loci. Although the number of statistical analysis and inference methods for Hi-C data is growing rapidly, a key impediment of available approaches is their inability to accommodate reads that align to multiple locations, i.e. multi-mapping reads. This is a key obstacle of current Hi-C pipelines and hinders the comprehensive investigation of both intra-chromosomal and inter-chromosomal interactions involving repetitive regions. We developed mHi-C, multi-mapping strategy for Hi-C data, integrates a hierarchical model to probabilistically allocate multi-mapping reads to their most likely genomic interaction positions. Application on published Hi-C data with varying sequencing depths demonstrates that a large fraction of novel significant contacts originates from heterochromatin regions of the genome, which were discarded in typical Hi-C pipeline due to their repetitive structure. Further analysis of these newly detected contacts for potential promoter-enhancer interactions highlights the importance of long-range contacts originating from duplicated segments. mHi-C is organized into a complete work-flow, starting from read alignment to significant contact detection, as a flexible and robust python pipeline which allows each main step to be run independently.

EO715 Room Gordon BIOINFORMATICS PIPELINES: FROM DATA COLLECTION TO STATISTICAL ANALYSIS Chair: Ekaterina Smirnova

E0538: **Sample size guided strategies for analysis of human microbiome data**
Presenter: **Alexander Alekseyenko**, Medical University of South Carolina, United States

The human microbiome datasets capture the abundances and compositions of entire microbial communities inhabiting anatomical microenvironments, obtained under a variety of study designs and sample size constraints. Pairwise distance matrices, known as beta diversity measures, are often the basis for many analyses in the microbiome space. We will introduce descriptive multivariate methods based on visualization of beta diversities. Significance of observed patters is often determined with PERMANOVA. We will present new versions of the PERMANOVA test, Tw2 and Wd*, which overcome its adverse behavior under heteroscedasticity and sample size imbalance. Likewise, we will present PERMANOVA-S method, which allows to draw inferences from microbiome data using ensembles of beta-diversity indices. The above techniques are useful when the sample size is small to medium, but have reduced relative utility with increasing sample size. When the sample size allows for robust inferences about individual microbes and sub-communities of microbes a different set of techniques allow for better inference. We will review approaches with medium to large sample sizes based on univariate testing with methods that link association with predictivity and causality. Results on the performance of feature selection and machine learning methods with microbiome data will be presented.

E0644: **ScAMP: Scalable automated metagenomics pipeline**
Presenter: **Lesley Hoyles**, Imperial College London, United Kingdom
Co-authors: James Abbott, Elaine Holmes, Jeremy Nicholson, Marc-Emmanuel Dumas, Sarah Butcher

An in-house pipeline was developed for the processing and analyses of sequence data generated from human microbiome studies. Quality analysis, trimming and filtering of sequence data allow reads associated with samples to be binned according to whether they represent human, prokaryotic, viral, parasite, fungal or plant DNA. Non-prokaryotic DNA can be assigned to species level on a presence/absence basis, allowing - for example - identification of dietary intake of plant-based foodstuffs and their derivatives. Prokaryotic DNA is subject to taxonomic analyses using MetaPhlAn2. After de novo assembly of sequence reads and gene prediction, a non-redundant catalogue of genes is built. From this catalogue, gene abundance and metagenomic species can be determined after normalization of data, as can microbial gene richness. Functional annotation of genes is achieved by mapping against KEGG proteins, InterProScan and CAZy. The pipeline was validated using previous data. Outputs from the pipeline allow development of tools for the integration of metagenomic and metabonomic data, moving metagenomic studies beyond determination of microbial gene richness and representation towards microbial-metabolite mapping.

E1214: **Assessing the ability of two recent algorithms to infer structure in longitudinal vaginal microbiome data**
Presenter: **Eugenie Jackson**, West Virginia University, United States

The study of microbial communities inhabiting human body sites has been an important area of research since their discovery. In the last 20 years, sequencing technologies have been developed that allow culture-free identification of community members. These communities are recognized for their important roles in the maintenance of good health and in the development of disease. Human Microbiome Projects 1 and 2, supported by the National Institutes of Health, have been instrumental in advancing these studies. The research presented here focuses on the bacterial communities specific to the human vagina, known as the vaginal microbiome. The data are characterized as sparse, high-dimensional, compositional, typically contaminated, and frequently involving more taxa than observations. These features necessitate the development of new methods for exploration and inference. The assessments of 2 promising algorithms are presented. BioMiCo (Bayesian inference of microbial communities), a supervised learning algorithm based on a hierarchical Bayesian model that infers an interpretable latent assemblage structure is considered first, followed by CORAL (Clustering and Ordination Regression AnaLysis), an unsupervised algorithm that classifies and clusters microbial data. Experiments designed to assess the capabilities and limitations of these algorithms when applied to longitudinal data are performed and recommendations are made.

E0806: **A robust and powerful statistical framework for differential abundance analysis of microbiome data**
Presenter: **Jun Chen**, Mayo Clinic, United States

One central theme of microbiome data analysis is to identify differentially abundant taxa. The identified taxa could provide insights into disease

etiology and, once validated, could be used as biomarkers for disease prevention and diagnosis. Due to the special characteristics of microbiome sequencing data, differential abundance analysis raises many statistical challenges, including modeling excessive zeros and overdispersion and taking into account the phylogenetic relationship among taxa. A robust and powerful framework is presented for differential abundance analysis of microbiome data. The framework consists of three parts (1) a fully generalized regression model based on the zero-inflated negative binomial model, which accounts for zero-inflation and overdispersion naturally and allows covariate-dependent dispersion to account for sample heterogeneity; (2) a new normalization method for zero-inflated sequencing data to address variable library sizes; and (3) a false discovery rate control procedure that integrates the phylogenetic tree to further improve the statistical power after differential abundance testing. The framework will be illustrated by using simulations as well as real data sets.

EO342 Room CLO 101 REGULARIZATION-/HYPER-/SMOOTHING-PARAMETERS SELECTION METHODS Chair: Sylvain Sardy

E0217: **Cross-validation for estimator selection**
Presenter: **Sylvain Arlot**, Universite Paris-Sud and INRIA, France
Co-authors: Alain Celisse, Matthieu Lerasle

Cross-validation is a widespread strategy because of its simplicity and its (apparent) universality. It can be used with two main goals: (i) estimating the risk of an estimator, and (ii) model selection or hyperparameter tuning, or more generally for choosing among a family of estimators. Many results exist on the performance of cross-validation procedures, which can strongly depend on the goal for which it is used. The big picture of these results will be shown, with an emphasis on the goal of estimator selection. In short, at first order (when the sample size goes to infinity), the key parameter is the bias of cross-validation, which only depends on the size of the training set. Nevertheless, second-order terms do matter in practice, and we will show recent results on the role of the "variance" of cross-validation procedures on their performance. As a conclusion, we will provide some guidelines for choosing the best cross-validation procedure according to the particular features of the problem at hand.

E1541: **Lasso-zero: Model selection by thresholding the ℓ_1-minimal solution**
Presenter: **Pascaline Descloux**, University of Geneva, Switzerland
Co-authors: Sylvain Sardy

The Lasso estimator is widely used in high-dimensional linear regression. However, no matter which criterion (cross-validation, SURE, BIC,) is used for selecting the regularization parameter λ, its performance in terms of model selection is limited: the proportion of false discoveries tends to be large and restrictive conditions are required on the design matrix for achieving exact recovery of the set of important variables. Rather than focusing on the first part of the Lasso path where sparsest solutions are obtained, it is suggested to consider the limiting solution as λ goes to zero, and to threshold the obtained coefficients. A choice of threshold aiming for low false discovery rate is proposed. The performance of this "Lasso-Zero" estimator is investigated and numerical experiments demonstrate that it provides an excellent tradeoff between false discovery and true positive rates and that it can exactly recover the set of important variables in situations where Lasso fails.

E1608: **Fast automatic smoothing in multiple generalized additive models**
Presenter: **Yousra El Bachir**, Ecole Polytechnique Fadarale de Lausanne, Switzerland
Co-authors: Anthony Davison

A general statistical methodology for fitting distributions with parameters that depend smoothly on covariates through additive structures is presented. These multiple generalized additive models (GAMs) are estimated simultaneously and the optimal degree of penalization, which determines their smoothness, is incorporated automatically through a likelihood-based approach. The resulting method is statistically efficient and numerically stable while being simpler and faster than the gold standard, and can be extended to big-data settings easily.

E1702: **Model selection as a multiple hypothesis testing procedure: Improving Akaike's information criterion**
Presenter: **Adrien Saumard**, Crest-Ensai, France
Co-authors: Fabien Navarro

By interpreting the model selection problem as a multiple hypothesis testing task in general, we propose a modification of Akaike's Information Criterion that avoids over-fitting. We call this correction an over-penalization of AIC. We prove nonasymptotic optimality of our procedure for histogram selection in density estimation, by establishing sharp oracle inequalities for the Kullback-Leibler divergence. A strong feature of our theoretical results is that they include the estimation of unbounded log-densities. To do so, we prove several analytical and probabilistic lemmas that are of independent interest. We also demonstrate the practical superiority of our over-penalization procedure over other model selection strategies in an extended, fully reproducible, experimental study. Our procedure is implemented in a R package.

EO402 Room CLO 102 SEMIPARAMETRIC TRANSFORMATION MODELS IN HEALTH AND SOCIAL SCIENCES Chair: Jan De Neve

E0415: **The proportional odds cumulative incidence model for competing risks**
Presenter: **Frank Eriksson**, University of Copenhagen, Denmark
Co-authors: Thomas Scheike, Li Jianing, Zhang Mei-Jie

An estimator for the proportional odds cumulative incidence model for competing risks data is suggested. The key advantage of this model is that the regression parameters have the simple and useful odds ratio interpretation. The model has been considered by many authors, but it is rarely used in practice due to the lack of reliable estimation procedures. Such procedures are suggested and it is shown that their performance improve considerably on existing methods. A goodness-of-fit test for the proportional odds assumption is also suggested. The large sample properties are derived and estimators of the asymptotic variance are provided. The method is illustrated by an application in a bone marrow transplant study.

E0338: **Probabilistic index models for flexible and efficient rank based inference**
Presenter: **Olivier Thas**, Ghent University, Belgium
Co-authors: Jan De Neve, Stijn Vansteelandt, Karel Vermeulen, Gustavo Amorim, Joris Meys

Probabilistic Index Models (PIM) are as a class of semiparametric models for the conditional probabilistic index (PI) which is defined as the probability that Y_2 exceeds Y_1 given X_1 and X_2, where X_1 and X_2 are the covariates corresponding to the outcomes Y_1 and Y_2, respectively. PIMs are related to semiparametric transformation models, but can be considered as more flexible since PIMs impose weaker restrictions on the conditional outcome distribution. The PIM methodology generates many of the classical rank tests for factorial designs, but the flexibility of the model allows to generate rank-type tests for many more complicated designs, including correcting for continuous covariate effects. The original PIM estimators were not efficient and simulation studies showed slow convergence. The focus is now on the construction of efficient estimators and on improved inference in small samples. A further development makes use of PIMs for increasing the power of the Wilcoxon test by using covariate information. This approach lends itself to permutation inference. We start with an introduction to the PIM with a focus on its flexibility for rank test generation, and on efficient parameter estimation and small sample inference. The relation with the Wilcoxon test and covariate adjustment will also be included. The methods are implemented in the PIM R package, which will be illustrated on several examples.

E0674: **Semiparametric analysis of two-sided markets**
Presenter: **Senay Sokullu**, University of Bristol, United Kingdom
An empirical semiparametric model is considered for two-sided markets. Contrary to existing empirical literature on two-sided markets, we specify network effects and probability distribution functions of net benefits of the two sides nonparametrically. We then estimate the model by nonparametric IV regression for local daily newspapers from the US. We show that semiparametric specification is supported by the data and the network effects are neither linear nor monotonic. With a numerical illustration we demonstrate that the mark-up of the newspaper on each side changes drastically with the nonlinearly specified network effects from the case with linear network effects.

E0334: **Semiparametric regression models for indirectly observed outcomes**
Presenter: **Jan De Neve**, Ghent University, Belgium
In several applications the outcome of interest is not measured directly, but instead a proxy is used. Examples include the body mass index as a proxy for body fat percentage, fluorescence intensity as a proxy for gene expression and the proportion of words correctly recalled as a proxy for the information stored in the memory. We illustrate by examples that the relationship between the outcome of interest and the proxy can be non-linear. The majority of the available methods, however, typically assume that this relationship is linear. Via simulations we illustrate how deviations from linearity can have a substantial impact on the validity of these procedures. We therefore present a semiparametric regression strategy to quantify the effect of covariates on a summary measure of the unobserved outcome, this without assuming linearity. We use the probabilistic index as a summary measure, i.e. the probability that the outcome of one subject exceeds the outcome of another subject, conditional on covariates. Since this effect measure is invariant under monotone transformations, we do not need to model the relationship between the unobserved outcome and the proxy. The estimation strategy makes use of semiparametric linear transformation models which enables us to use existing software packages for data analysis.

EO184 Room Jessel MODEL ASSESSMENT Chair: Maria Dolores Jimenez-Gamero

E0274: **Mean integrated squared error comparison between kernel and maximum likelihood density estimates of normal mixtures**
Presenter: **Dimitrios Bagkavos**, University of Crete, Greece
Co-authors: Prakash Patil
A methodological advance is presented in order to help practitioners to decide in selecting between parametric and nonparametric estimates when estimating mixtures of normal distributions. Through graphical tools and simulations numerical evidence is provided indicating that, as expected, the parametric approach is a more accurate choice when the number of mixture components is small. As the number of components increase, the difficulty in precisely estimating their number, together with the convergence issues that might occur in the implementation of the maximum likelihood algorithm, point to the direction of kernel based estimates as a more reliable option. Further investigation sheds light on the choice between simple and the variable bandwidth version of kernel density estimates.

E0832: **Penalized minimum phi-divergence estimators in multinomial models**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Maria Dolores Jimenez-Gamero
The aim is to study the consequences of model misspecification for multinomial data when using penalized minimum Phi-divergence or penalized minimum disparity estimators to estimate the model parameters. These estimators are shown to converge to a well-defined limit. As an application of the results obtained, we consider the problem of testing goodness-of-fit to a given parametric family for multinomial data, using as test statistic a penalized divergence between the observed frequencies and a estimation of the null model cell probabilities. In some previous simulation studies, it has been observed that the asymptotic approximation to the null distribution of the test statistics in this class is rather poor. As an alternative way to approximate this null distribution, we prove that the bootstrap consistently estimates it. We present a numerical example illustrating the convenience of the bootstrap approximation which, in spite of demanding more computing time, it is more accurate than the approximation yielded by the asymptotic null distribution.

E0817: **Goodness-of-tit tests for GARCH models**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidade de Vigo, Spain
Co-authors: Simos Meintanis, Sangyeol Lee
Goodness-of-fit tests for correct specification of GARCH type models are proposed. This problem has been previously considered by other authors. Many of the proposed methods suffer from at least one of the following drawbacks: (i) being specific to a particular GARCH model, often of given (low) order, (ii) assume rather strong conditions (such as the existence of certain moments), and (iii) not being consistent under all fixed alternatives. On the contrary, our method is for arbitrary specification of the model; it will be seen to be consistent under arbitrary model deviations.

EO212 Room Montague STATISTICAL BOOSTING Chair: Andreas Mayr

E1008: **Tuning model-based gradient boosting algorithms with focus on variable selection**
Presenter: **Tobias Hepp**, Friedrich-Alexander-Universitaet Erlangen-Nuernberg, Germany
Co-authors: Janek Thomas, Andreas Mayr, Bernd Bischl
Variable selection in regularized regression models like the lasso or gradient boosting algorithms is usually controlled by method-specific tuning-parameters that define the degree of penalization. While these parameters are commonly determined using resampling strategies like cross-validation, bootstrapping and similar methods, their focus on minimizing the prediction error often results in the selection of many variables without true effect on the outcome. Therefore, we propose a new method to determine the optimal number of iterations in model-based boosting for variable selection inspired by probing, a method used in related areas of machine learning research. The general notion of probing involves the artificial inflation of the data with random noise variables, so-called probes or shadow variables. Using the first selection of a shadow variable as stopping criterion, the algorithm is applied only once without the need to optimize any hyperparameters in order to extract a set of informative variables from the data, thereby making its application very fast and simple in practice. Furthermore, simulation studies show that the resulting models tend to be more strictly regularized compared to the ones resulting from cross-validation, thereby substantially reducing the high number of false discoveries.

E0513: **Selective inference for boosting**
Presenter: **David Ruegamer**, LMU Munich, Germany
Co-authors: Sonja Greven
The necessity for an explicit inference framework after model selection is due to the invalidity of classical inference after model selection. Following recent work on this topic, we address the issue of conducting valid inference when using boosting for variable selection and model fitting. We therefore make use of a recently proposed inferential framework called selective inference, which corrects the inference after the selection procedure. In detail, we apply the polyhedron approach to L_2-boosting in order to perform inference for boosted linear models. As the polyhedron approach conditions on most of the information in the data, thereby leaving no information for a powerful inference post selection, we motivate the

use of an adaption of the approach which is a powerful and more coherent way to perform selective inference for L_2-boosting. By circumventing an explicit mathematical definition of the selective inference space, our algorithm is a flexible tool, which is usually less expensive than Bootstrap and may also be used to incorporate resampling schemes or stability selection.

E0599: **Statistical boosting in Markov-switching generalized additive models for location, scale and shape**
Presenter: **Timo Adam**, Bielefeld University, Germany
Co-authors: Andreas Mayr, Thomas Kneib

A novel class of flexible latent-state time series regression models is proposed which is called Markov-switching generalized additive models for location, scale and shape. In contrast to conventional Markov-switching regression models, the presented methodology allows us to model state-dependent parameters of the response beyond the mean - including variance, skewness and kurtosis parameters - as potentially smooth functions of a given set of explanatory variables. In addition, the set of possible distributions that can be specified for the response is not limited to the exponential family but additionally includes, for instance, a variety of Box-Cox-transformed, zero-inflated and mixture distributions. We derive a novel EM algorithm and demonstrate how statistical boosting can be exploited to prevent overfitting while simultaneously performing variable selection. The suggested approach is assessed in simulation experiments and illustrated in a real-data setting, where we model the conditional distribution of the daily average price of energy in Spain over time.

E0601: **Regularized censored regression with conditional heteroskedasticity**
Presenter: **Jakob Wolfgang Messner**, Technical University of Denmark, Denmark
Co-authors: Achim Zeileis

Censored regression models such as the classic tobit model are often used to model limited responses, typically non-negative variables censored at zero. While originally employed for economic applications, e.g., the expenditures for durable goods, such models recently gained popularity in the atmospheric sciences for modeling precipitation or wind power. In this context it is important not only to link the mean to the regressors, but also include a submodel for the scale of the response distribution. When applying this heteroskedastic censored regression to high-dimensional data, over-fitting can occur and decrease the predictive performance.To avoid this problem we present two methods: a gradient boosting approach and a coordinate descent algorithm to derive lasso paths. Both methods regularize the coefficients and can be used to automatically select the most relevant input variables. We test and compare these algorithms on weather data to derive probabilistic precipitation forecasts on the basis of various outputs of numerical weather prediction models.

EO160 Room Senate ADVANCES IN ROBUST DATA ANALYSIS Chair: Luis Angel Garcia-Escudero

E1391: **Visual tools for 3-way analysis in R**
Presenter: **Valentin Todorov**, UNIDO, Austria
Co-authors: Michele Gallo, Maria Anna Di Palma

The standard multivariate analysis addresses data sets represented as two dimensional matrices. In recent years, an increasing number of application areas like chemometrics, computer vision, econometrics and social network analysis involve analysis of data sets that are represented as multidimensional arrays and multi-way data analysis becomes popular as an exploratory analysis tool. The most popular trilinear models are PARAFAC and Tuccker3 and their results can be presented in several different ways, the first one being tables of the coefficients or loadings for each mode, either rotated or not. While it is important to inspect the numerical output of the methods for analysis of three-way data in order to properly interpret the results, of great help can be different visual representations of these outcomes. We present an R package, rrcov3way, implementing a set of functions for the analysis of three-way data sets, including PARAFAC and Tucker3 as well as their robust alternatives. Apart from basic tools for data handling and preprocessing of multidimensional arrays, tools for display of the raw data and the model results in two and three dimensional plots are provided.

E1734: **A robust clustering procedure with unknown number of clusters**
Presenter: **Francesco Dotto**, Sapienza - University of Rome, Italy
Co-authors: Alessio Farcomeni

A new methodology for robust clustering without specifying in advance the underlying number of Gaussian clusters is proposed. The procedure is based on iteratively trimming, assessing the goodness of fit, and reweighting. The forward version of our procedure proceeds by fixing a high trimming level and $K = 1$ population. The procedure is then iterated throughout a fixed sequence ofdecreasing trimming levels. New observations are added at each step and, whenever necessary, the number of components K is increased. Goodness of fit is assessed against the empirical distribution of theMahalanobis distances of the untrimmed observations from the closest centroid, with parameters estimated at the previousiteration. A stopping rule prevents our procedure for using outlying observations; while a backward criterion is adopted whenever too many clusters are detected. A simulation study shows that our method compares well with robust procedures with known number of clusters, and is robust in the presence of different contamination schemes.

E1355: **Outlier detection in multivariate data with robust Mahalanobis distance based on shrinkage estimators**
Presenter: **Elisa Cabana**, University Carlos III of Madrid, Spain
Co-authors: Henry Laniado Rodas, Rosa Lillo

Different combinations of robust location and covariance matrix estimators based on the notion of Shrinkage are proposed. These collection defines robust Mahalanobis distances to address the problem of detecting outliers in multivariate data. The parameters needed for the shrinkage estimators defined, have been optimally estimated. The performance of the proposed distances is studied by means of a comparison with other existing methods from the literature, in simulated scenarios and with a real dataset example. The good computational results and the high correct detection rates and low false detection rates in the vast majority of cases, shows the advantages of our proposal.

E1839: **The power of monitoring: How to make the most of a contaminated multivariate sample**
Presenter: **Marco Riani**, University of Parma, Italy
Co-authors: Anthony Atkinson, Andrea Cerioli, Aldo Corbellini

Diagnostic tools must rely on robust high-breakdown methodologies to avoid distortion in presence of contamination by outliers. However, a disadvantage of having a single, even if robust, summary of the data is that important choices have to be made prior to the analysis and their effect may be difficult to evaluate. We argue that an effective solution is to look at several pictures, and possibly to a whole movie, of the available data. This is what we obtain by monitoring the results computed through the robust methodology of choice. We show the information gain that monitoring provides in the study of complex data structures through the analysis of multivariate datasets and using different high-breakdown techniques. Our findings support the claim that the principle of monitoring is very flexible and that it can lead to robust estimators that are as efficient as possible. We also address through simulation some of the tricky inferential issues that arise from monitoring.

EO396 Room Woburn STATISTICAL INFERENCE IN NETWORKS Chair: Tengyao Wang

E0782: **A semidefinite program for structured blockmodels**
Presenter: **David Choi**, Carnegie Mellon, United States
Semidefinite programs (SDP) have been intensely studied for a particular class of the stochastic blockmodel, which exhibits assortative connectivity and corresponds to community detection. However, there exist blockmodels outside of this class for which the known SDP formulation is not applicable, and it is of interest to consider whether this is an inherent limitation to the semidefinite programming approach, or if alternate formulations exist. We present a family of semidefinite programs that can be tailored to such instances of the blockmodel, such as non-assortative networks and overlapping communities. We establish label recovery in sparse settings, with conditions that are analogous to known (though not the best known) results for community detection. When the blockmodel exhibits symmetry or label-switching ambiguities, the computation time the SDP can be significantly reduced by parameterizing out the non-identifiable subspace, using a concept known in combinatorics as an association scheme.

E0810: **Maximum likelihood estimation of preferential attachment network models**
Presenter: **Fengnan Gao**, Fudan University and SCMS, China
The preferential attachment (PA) network is a popular way of modeling the social networks, the collaboration networks and etc. The PA network model is an evolving network model where new nodes keep coming in. When a new node comes in, it establishes only one connection with an existing node. The random choice on the existing node is via a multinomial distribution with probability weights based on a preferential function f on the degrees. f maps the natural numbers to the positive real line and is assumed a priori non-decreasing, which means the nodes with high degrees are more likely to get new connections, i.e. "the rich get richer". If f is affine with $f(k) = k + \delta$, it is well known that such a model leads to a power-law degree distribution. We proposed a maximum likelihood estimator for delta and establish a central limit result on the MLE of delta. If f belongs to a parametric family no faster than linear, we show the MLE will also yield optimal performance with the asymptotic normality results. We will also consider the potential extensions of the model (with borrowed strength from nonparametric Bayesian statistics) and interesting applications.

E1269: **Sampling on social networks from a decision theoretic perspective**
Presenter: **Simon Lunagomez**, Lancaster University, United Kingdom
Co-authors: Edoardo Airoldi
Some of the most used sampling mechanisms that propagate through a social network are defined in terms of tuning parameters; for instance, Respondent Driven Sampling (RDS) is specified by the number of seeds and maximum number of referrals. We are interested in the problem of optimising these tuning parameters with the purpose of improving the inference of a population quantity, where such quantity is a function of the network and measurements taken at the nodes. This is done by formulating the problem in terms of Decision Theory. The optimisation procedure for different sampling mechanisms is illustrated via simulations in the fashion of the ones used for Bayesian clinical trials.

E1240: **Optimal rates for community estimation in the weighted stochastic block model**
Presenter: **Min Xu**, University of Pennsylvania, United States
Co-authors: Po-Ling Loh, Varun Jog
Community identification in a network is an important problem in fields such as social science, neuroscience, and genetics. Over the past decade, stochastic block models (SBMs) have emerged as a popular statistical framework for this problem. However, SBMs have an important limitation in that they are suited only for networks with unweighted edges; in various scientific applications, disregarding the edge weights may result in a loss of valuable information. We study a weighted generalization of the SBM, in which observations are collected in the form of a weighted adjacency matrix and the weight of each edge is generated independently from an unknown probability density determined by the community membership of its endpoints. We characterize the optimal rate of misclustering error of the weighted SBM in terms of the Renyi divergence of order 1/2 between the weight distributions of within-community and between-community edges, substantially generalizing existing results for unweighted SBMs. Furthermore, we present a principled, computationally tractable algorithm based on discretization that achieves the optimal error rate without assuming knowledge of the weight densities.

EO549 Room CLO 203 STATISTICS FOR DATA WITH GEOMETRIC STRUCTURE Chair: Stephan Huckemann

E0484: **Small-sphere distributions for directional data with application to rotationally deformed objects**
Presenter: **Joern Schulz**, University of Stavanger, Norway
Co-authors: Byungwon Kim, Stephan Huckemann, Sungkyu Jung
Rotational deformations of 3D objects such as bending or twisting have been observed as the major variations in various medical applications. To provide a better surgery or treatment planning, it is crucial to model such deformations that can be described by the movements of multivariate directional vectors on $(S^2)^K$ for $K \geq 1$. Such multivariate directional vectors are available in a number of different object representations. The proposed small-sphere distribution families enable to model these directions and their dependence structure. In addition, they facilitate random data generation and hypotheses testing of the directional data. A Likelihood-based estimation procedure is suggested for the estimation of the corresponding parameters. The proposed models will be compared to a non-parametric approach where estimates of the rotation axis of small-circle-concentrated data are obtained by fitting small circles applying sample Frechet means and least-square estimators. The performance of the proposed multivariate small-sphere distributions is demonstrated: i.) in a simulation study, ii.) on deformed ellipsoids and iii.) on knee motions during gait.

E1262: **Maximum likelihood estimation for general models in size and shape space**
Presenter: **Alfred Kume**, University of Kent, United Kingdom
Co-authors: Ian Dryden, Phillip Paine, Andrew Wood
Inference in shape analysis is related to problems where the invariance of rotations and translations (and possibly scaling) of data objects is required. If some multivariate distribution is assumed in the landmarks space the estimation is not straightforward. We will represent a general approach for maximum likelihood estimation for such models in size-and-shape spaces, where a general time dependence and covariance among landmarks is allowed. In particular, while for the 2-d case we can use the Bessel functions to carry out the MLE procedure, in the 3-d case we make use of the recent numerical methods for calculating the quantities of interest. Real and simulated data are used to illustrate the proposed method.

E1265: **Geometric statistics for template shape estimation in computational anatomy**
Presenter: **Nina Miolane**, Inria Stanford, France
Statistics is a science that studies methods of inference from observed data which often belong to vector spaces i.e. linear spaces. In contrast, Geometric statistics generalizes statistics for data belonging to manifolds, i.e. non-linear spaces. Such non-linear data spaces emerge naturally in Computational Anatomy. Organ shapes can be modeled as the equivalence of their configurations in the 3D space under the action of rotations and translations. In this case, they are elements of a quotient space, which is a stratified manifold. Geometric Statistics on stratified manifolds allow to

analyze the properties of the algorithm of template (organ) shape computation. We show that this algorithm - used for more than 15 years in the medical imaging community - has a systematic bias. We also provide correction methods.

E0970: **Inferential PCA and RNA structure analysis**
Presenter: **Benjamin Eltzner**, Georg-August-University of Goettingen, Germany
Co-authors: Stephan Huckemann
RNA backbone geometry is canonically described in terms of dihedral angles, which are represented on a torus. PCA on a torus is challenging and so is variance analysis and dimension reduction: There is always a price to pay. We pay a small price and develop an asymptotic theory for our approach that gives rise to two-sample tests via bootstrapping.

EO318 Room CLO 204 OPTIMAL AND EFFICIENT DESIGNS Chair: Chang-Yun Lin

E0726: **Robust split-plot designs for model misspecification**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan
Many existing methods for constructing optimal split-plot designs, such as D-optimal or A-optimal designs, only focus on minimizing the variance of the parameter estimates for the fitted model. However, the true model is usually more complicated and, hence, the fitted model is often misspecified. If there exist significant effects that are not included in the model, then the estimates could be highly biased. Therefore, a good split-plot design should be able to simultaneously control the variance and the bias of the estimates. We propose a new method for constructing optimal split-plot designs that are robust under model misspecification. Four examples are provided to demonstrate that our method can produce efficient split-plot designs with smaller overall aliasing. Simulation studies are performed to verify that the robust designs we construct have high power, low false discovery rate, and small mean squared error.

E1199: **Sequential integer quadratic programming for generalized minimum aberration orthogonal arrays**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy
Orthogonal Fractional Factorial Designs and in particular Orthogonal Arrays are frequently used in many fields of application, including medicine, engineering and agriculture. We present a methodology and an algorithm to find an orthogonal array, of given size and strength, that satisfies the generalized minimum aberration criterion. The methodology is based on the joint use of polynomial counting functions, complex coding of levels and algorithms for quadratic optimization and puts no restriction on the number of levels of each factor.

E0254: **Optimal allocation for response-adaptive designs**
Presenter: **Yanqing Yi**, Memorial University of Newfoundland, Canada
A new allocation proportion is derived by using differential equation methods for response-adaptive designs. This new allocation is compared with the balanced and the Neyman allocations and other optimal allocations from an ethical point of view and statistical power performance. Theoretically, the new allocation is proven to have the ethical advantages of allocating more than 50% of patients to the better treatment. It also allocates higher proportion of patients to the better treatment than the existing adaptive optimal allocation for success probabilities larger than 0.5. The statistical power under the proposed allocation is compared with the existing allocations through simulation. The simulation results indicate that the statistical power under the proposed allocation proportion is similar as to those under the balanced, the Neyman and the existing adaptive optimal allocations.

E1447: **Mixture-process designs and response surface models for predicting properties of calcium silicate units**
Presenter: **Sonja Kuhnt**, Dortmund University of Applied Sciences and Arts, Germany
Calcium silicate units ensure good acoustic and thermal insulation as well as fire resistance when used for building houses. In the production process a mixture of sand, lime and water is high-pressed first and then autoclaved. We present an experiment with mixture components and controllable process variables. The main interest lies in studying the compressive strength of the produced calcium silicate units. Different D- and I-optimal designs are considered and compared by their efficiencies and prediction variance. We discuss specific response surface models for mixture-process experiments. Results from running the chosen design are analysed and their use in predicting compressive strength is shown.

EO623 Room SH349 RECENT DEVELOPMENTS IN MIXTURE MODELS Chair: Salvatore Ingrassia

E1110: **Finite mixture modeling of censored data using the multivariate Student-t distribution**
Presenter: **Victor Hugo Lachos Davila**, University of Connecticut, United States
Finite mixture models have been widely used for the modeling and analysis of data from a heterogeneous population. Moreover, data of this kind can be subject to some upper and/or lower detection limits because of the restriction of experimental apparatus. Another complication arises when measures of each population depart significantly from normality, for instance, in the presence of heavy tails or atypical observations. For such data structures, we propose a robust model for censored data based on finite mixtures of multivariate Student-t distributions. This approach allows us to model data with great flexibility, accommodating multimodality, heavy tails and also skewness depending on the structure of the mixture components. We develop an analytically simple, yet efficient, EM-type algorithm for conducting maximum likelihood estimation of the parameters. The algorithm has closed-form expressions at the E-step that rely on formulas for the mean and variance of the multivariate truncated Student-t distributions. Further, a general information-based method for approximating the asymptotic covariance matrix of the estimators is also presented. Results obtained from the analysis of both simulated and real data sets are reported to demonstrate the effectiveness of the proposed methodology. The proposed algorithm and methods are implemented in the new R package CensMixReg.

E1027: **Model based clustering via pair copula and applications**
Presenter: **Marta Nai Ruscone**, LIUC, Italy
Finite mixtures are applied to perform model-based clustering of multivariate data. Existing models do not offer great flexibility for modelling the dependence of multivariate data since they rely on potentially undesirable correlation restrictions to be computationally tractable. We propose a model-based clustering method via pair copula to reveal and fully understand the complex and hidden dependence patterns in correlated multivariate data. Since this approach is based on pair copula constructions it takes into account also tail asymmetry of the data by using blocks of asymmetric bivariate copulas.We use simulated and real datasets to illustrate the proposed procedure.

E1076: **Model-based clustering of variables**
Presenter: **Vincent Vandewalle**, Inria, France
Co-authors: Thierry Mottet
In the clustering of variables framework, the goal is to cluster together similar variables based on a distance between variables. This distance can be easily defined when dealing with variables of the same type, but is more difficult to define when dealing with variables of different types. In this communication we propose a model-based clustering of variables. It consists in grouping together variables defining the same groups of individuals. In variables of the same cluster, a conditional independence model is assumed for the clustering of the individuals. This model has the advantage of only needing to define probability distribution functions from the univariate point of view, and allows a simple clustering of

variables (one clustering per cluster of variables) when partitions of individuals are known. The clustering of variables becomes a model selection issue which is answered by optimizing the BIC criterion through a modified version of the EM algorithm. The proposed approach is illustrated on simulated and real data.

E0578: **Clustering airbnb reviews**
Presenter: **Yang Tang**, McMaster University, Canada
Co-authors: Paul McNicholas
A clustering approach is developed for Boston Airbnb reviews, in the English language, collected since 2008. This approach is based on a mixture of latent variables model, which provides an appealing framework for handling clustered binary data. In the broader context of social science applications (e.g., voting data, web reviews, and survey data), extremely large numbers of variables rule out the use of a mixture of latent trait models. A penalized mixture of latent traits approach is developed to reduce the number of parameters and identify variables that are not informative for clustering. The introduction of component-specific rate parameters avoids the over-penalization that can occur when inferring a shared rate parameter on clustered data. A variational expectation-maximization algorithm is developed and provides closed-form estimates for model parameters; this is in contrast to an intensive search over the rate parameters via a model selection criterion. This approach is important for a whole class of applications, but the focus herein is the Boston Airbnb reviews data.

CO459 Room MAL B35 TOPICS IN MACROECONOMETRICS Chair: Alessia Paccagnini

C1025: **News and uncertainty shocks**
Presenter: **Danilo Cascaldi-Garcia**, University of Warwick, United Kingdom
Co-authors: Ana Galvao
Novel empirical evidence linking the effects of technology news shocks with uncertainty shocks is provided. Their observed linkage implies that when financial uncertainty shocks hit the economy, utilization-adjusted total factor productivity increases over the medium term. This leads to an attenuation of the effects on economic activity from technology news shocks in the short term and from uncertainty shocks in the medium term. Supported by these results, we propose an identification strategy to measure the effects of 'good uncertainty' shocks and disentangle the importance of technological news, good and bad uncertainties, and ambiguity shocks in explaining business cycle variation.

C0776: **Uncertainty shocks and monetary policies**
Presenter: **Valentina Colombo**, University of Bologna, Italy
The effects of financial uncertainty shocks on monetary policy in the United States are investigated. Financial uncertainty is captured by appealing to some indicators recently developed. Relying on a nonlinear VAR, we isolate the effects of uncertainty shocks in both recessionary and expansionary periods. Uncertainty shocks trigger negative macroeconomic fluctuations across the business cycle. The asymmetric effects are persistent when we include the recent recession. To reduce the fall in macroeconomic variables, the unconventional monetary policy plays an important role.

C0570: **The federal reserve's implicit inflation target and macroeconomic dynamics: A SVAR analysis**
Presenter: **Haroon Mumtaz**, Queen Mary University of London, United Kingdom
Co-authors: Theodoridis Konstantinos
The purpose is to identify shocks to the Federal Reserve's inflation target as VAR innovations that make the largest contribution to future movements in long-horizon inflation expectations. The effectiveness of this scheme is documented via Monte-Carlo experiments. The estimated impulse responses indicate that a positive shock to the target is associated with a large increase in inflation and long-term interest rates in the US and the industrialised world. Target shocks are estimated to be a vital factor behind the increase in inflation during the pre-1980 period and are an important driver of the decline in long-term interest rates over the last two decades.

C1095: **Forecasting with FAVAR: Macroeconomic versus financial factors**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland
The purpose is to assess the predictive power of macroeconomic and financial latent factors on the key variables for the US economy before and after the recent Great Recession. We implement a forecasting horse race among Factor Augmented VAR(FAVAR), Classical, and Bayesian VAR models. FAVAR models outperform others. Focusing only on macroeconomic or on financial latent factors, we find how the financial variables have not a driver role in forecasting the US economy including the Great Recession.

CO742 Room MAL 402 CENTRAL BANK FORECASTING II Chair: Knut Are Aastveit

C0774: **Mixed frequency models with MA components**
Presenter: **Claudia Foroni**, Deutsche Bundesbank, Germany
Co-authors: Massimiliano Marcellino, Dalibor Stevanovic
Temporal aggregation in general introduces a moving average (MA) component in the aggregated model. A similar feature emerges when not all but only a few variables are aggregated, which generates a mixed frequency model. The MA component is generally neglected, likely to preserve the possibility of OLS estimation, but the consequences have never been properly studied in the mixed frequency context. We show, analytically, in Monte Carlo simulations and in a forecasting application on U.S. macroeconomic variables, the relevance of considering the MA component in mixed-frequency MIDAS and UMIDAS models (MIDAS-ARMA and UMIDAS-ARMA). Specifically, the simulation results indicate that the short-term forecasting performance of MIDAS-ARMA and UMIDAS-ARMA is better than that of, respectively, MIDAS and UMIDAS, and the empirical applications confirm this ranking.

C1183: **Nowcasting earnings**
Presenter: **Michele Modugno**, Federal Reserve Board, United States
Like GDP, companies earnings are quarterly data that begin to be released one or two weeks after the last month of each quarter. In other words, the majority of the public companies releases their earnings in early to mid- January, April, July and October. We propose a model to nowcast S&P500 companies earnings, that makes use of data at higher frequency, and that are published in a more timely manner than earnings. Nowcasting earnings is extremely important given that they are the primary metric to evaluate the performance of a company. Moreover, market participants form expectations about them, and every time earnings are released, and they differ from market expectations stock price of publicly traded company move dramatically. We find that in order to predict the average earnings of S&P500 companies a factor that capture the real side of the U.S. economy is enough to generate nowcast that are more accurate than market expectations, while factors capturing inflation, monetary policy, and the same stocks return have limited impact on improving the accuracy of the nowcast.

C1206: **Combining long-run survey forecasts and nowcasts with VAR forecasts using relative entropy**
Presenter: **Ellis Tallman**, Federal Reserve Bank of Cleveland, United States
Co-authors: Saeed Zaman, Ellis Tallman

Previous research highlights how nowcasts can improve the forecast accuracy in both Vector Autoregressions (VARs) and Dynamic Stochastic General Equilibrium (DSGE) models. Research has also highlighted the superior performance of the long-horizon survey forecasts compared to econometric approaches to forecasting long-horizon trajectories. We use real-time forecast evaluation to show that combining both long and short term conditions within one modeling approach generates meaningful gains in forecast accuracy. Specifically, we combine VAR forecasts with both external nowcasts and long-horizon survey forecasts using relative entropy to refine the medium-term forecasts of the VAR. The horizon at which we combine the VAR forecast with the long-horizon survey forecast varies by variable depending upon the degree of persistence of the variable. We propose a simple method to determine the relevant horizon for combination. Given that surveys are performed infrequently and do not cover all the forecast horizons, our procedure can also be thought as an approach to interpolate survey forecasts.

C0948: **Multivariate Bayesian predictive synthesis in macroeconomic forecasting**
Presenter: **Knut Are Aastveit**, Norges Bank, Norway
Co-authors: Kenichiro McAlinn, Jouchi Nakajima, Mike West

The aim is to develop the methodology of Bayesian predictive synthesis (BPS) in multivariate time series forecasting with a detailed application in multi-step macro-economic forecasting. Based on foundations in coherent Bayesian reasoning with predictive and decision analytic goals, BPS defines a methodological framework for evaluation, calibration, comparison, and context- and data-informed combination of multiple forecast densities. The BPS framework naturally allows modeling and estimation– sequentially and adaptively over time– of varying forecast biases and facets of miscalibration of individual forecast densities, and critically of time-varying inter-dependencies among models or forecasters over multiple series.BPS analysis is developed in one subset of the implied dynamic multivariate latent factor model. Bayesian simulation-based computation enables implementation. A study of sequential BPS analysis in a multiple macroeconomic time series study with US data highlights the potential improvement of forecasts among all series and forecast horizons, as well as the dynamic relationships among forecasting agents over multiple series.

CO506 Room MAL 414 ADVANCES IN DYNAMIC MACROECONOMIC MODELING Chair: Matteo Fragetta

C1270: **The government spending multiplier at the zero lower bound: Evidence from the United States**
Presenter: **Matteo Fragetta**, University of Salerno, Italy
Co-authors: Emanuel Gasteiger, Mario Di Serio

How large is the government spending multiplier in normal times and how large is it when monetary policy is constrained by the zero interest rate lower bound (ZLB)? It is frequently argued that in such a situation an increase in government spending is even more effective. A growing theoretical literature examines this claim. Likewise, there is an emerging literature developing reasonable theories that suggest that the government spending multiplier at the ZLB is 1 or below, and lower than in times without the ZLB binding. The objective is to provide further state-dependent evidence from the United States on the size of the aggregate government spending multiplier at the ZLB. Independent from the specification, our results show that multipliers at the ZLB always exceed their counterparts away from the ZLB. Thus, our results are qualitatively and quantitatively consistent with the claim that increases in government spending are even more effective at the ZLB.

C0772: **Tractable estimation and smoothing of highly non-linear dynamic state-space models**
Presenter: **Tom Holden**, University of Surrey, United Kingdom

A method is presented for tractably estimating large, non-linear dynamic state-space models, such as multi-asset stochastic volatility models, or DSGE models with occasionally binding constraints, based on approximating their likelihood. The method approximates the distribution of unknown states by an extended skew-t distribution, allowing it to track their first four moments without sacrificing the computational and econometric advantages of a parametric approach. We show that the extended skew-t distribution maintains the properties of the Gaussian distribution that gave tractability to the Cubature Kalman Filter (CKF). Our method extends the CKF by introducing alternative cubature procedures, to further improve the tracking of non-linearities, an augmented-state representation, reducing integration requirements, and dynamic state space reduction, to ensure that it can handle the large state spaces generated, for example, by pruned perturbation solutions to medium-scale DSGE models. We also show how a modified objective can deliver efficient, consistent estimates despite the approximations inherent in our method.

C0672: **Nonlocal solutions to dynamic equilibrium models: The approximate stable manifolds approach**
Presenter: **Viktors Ajevskis**, Bank of Latvia and Riga Technical University, Latvia

A method for constructing a sequence of approximate solutions of increasing accuracy to general equilibrium models on nonlocal domains is presented. The method is based on a technique originated from dynamical systems theory. The approximate solutions are constructed employing the contraction mapping theorem and the fact that solutions to general equilibrium models converge to a steady state. Under certain nonlocal conditions the convergence of the approximate solutions to the true solution is proved. The proposed approach can be treated as a rigorous proof of convergence for the extended path algorithm a class of nonlinear rational expectation models.

C0311: **On nonlinearities in unemployment**
Presenter: **Francois Langot**, Le Mans University, France
Co-authors: Frederic Karame, Stephane Adjemian, Frederic Karame

An extended version of the Diamond-Mortensen-Pissarides model is proposed to explain the nonlinearities of the US labor market. We take into account occasionally binding constraints and integrate worker heterogeneity, endogenous firing costs and a minimum wage. Four regimes can be distinguished: in the first, it is optimal to hire; in the second it is optimal to do nothing; in the third it is optimal to fire workers; and the last one it is optimal to close the firm. Aggregate dynamics combine heterogeneous reactions, specific to each labor market segmented by the ability types. Combining global methods to solve the model and particle filtering to estimate the structural parameters, take advantage of the information on nonlinearities that are contained in the data. We show that it provides a very good fit of the three time series of interest, the job finding, the job separation and the unemployment rates. The estimation shows that the firing costs are insignificant. These extensions of the basic DMP model allow us to investigate the cyclical behaviors of the vacancy rate and of the wage distribution, as well as the impact of a minimum wage increase conditionally to a business cycle episode.

CO124 Room MAL 415 ADVANCES IN VOLATILITY MODELLING Chair: Genaro Sucarrat

C0291: **Exponential-type GARCH models with linear-in-variance risk premium**
Presenter: **Dimitra Kyriakopoulou**, Universite Catholique de Louvain, Belgium
Co-authors: Christian Hafner
One of the implications of the intertemporal capital asset pricing model (CAPM) is that the risk premium of the market portfolio is a linear function of its variance. Yet, estimation theory of classical GARCH-in-mean models with linear-in-variance risk premium requires strong assumptions and is incomplete. We show that exponential-type GARCH models such as EGARCH or Log-GARCH are more natural in dealing with linear-in-variance risk premia. For the popular and more difficult case of EGARCH-in-mean, we derive conditions for the existence of a unique stationary and ergodic solution and invertibility following a stochastic recurrence equation approach. We then show consistency and asymptotic normality of the quasi maximum likelihood estimator under weak moment assumptions. An empirical application estimates the dynamic risk premia of a variety of stock indices using both EGARCH-M and Log-GARCH-M models.

C0647: **A new approach to volatility modeling: The factorial hidden Markov volatility model**
Presenter: **Arnaud Dufays**, Laval University, Canada
Co-authors: Maciej Augustyniak, Luc Bauwens
A new model – the factorial hidden Markov volatility (FHMV) model – is proposed for financial returns and their latent variances. It is also applicable to model directly realized variances. Volatility is modeled as a product of three components: a Markov chain driving volatility persistence, an independent discrete process capable of generating jumps in the volatility, and a predictable (data-driven) process capturing the leverage effect. The Markov chain and jump components allow volatility to switch abruptly between thousands of states. The transition probability matrix of the Markov chain is structured in such a way that the multiplicity of the second largest eigenvalue can be greater than one. This distinctive feature generates a high degree of volatility persistence. The statistical properties of the FHMV model are derived and an economic interpretation is attached to each component. In-sample results on six financial time series highlight that the FHMV model compares favorably to the main existing volatility processes. A forecasting experiment shows that the FHMV model significantly outperforms its competitors when predicting volatility over time horizons longer than five days.

C0408: **Testing normality for unconditionally heteroscedastic macroeconomic variables**
Presenter: **Hamdi Raissi**, PUCV, Chile
Testing normality for unconditionally heteroscedastic macroeconomic time series is considered. It is underlined that the classical Jarque-Bera test for normality is inadequate in our framework. On the other hand, it is found that the approach which consists in correcting the heteroscedasticity by kernel smoothing for testing normality is justified asymptotically. Nevertheless, it appears from Monte Carlo experiments that such a methodology can noticeably suffer from size distortion for samples that are typical for macroeconomic variables. As a consequence, a bootstrap methodology for correcting the problem is proposed. The innovations distribution of a set of inflation measures for the U.S., Korea and Australia are analyzed.

C0403: **Equation-by-equation estimation of multivariate periodic electricity price volatility**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway
Electricity prices are characterised by strong autoregressive persistence, periodicity (e.g. intraday, day-of-the week and month-of-the-year effects), large spikes or jumps, GARCH and - as evidenced by recent findings - periodic volatility. We propose a multivariate model of volatility that decomposes volatility multiplicatively into a non-stationary (e.g. periodic) part and a stationary part with log-GARCH dynamics. Since the model belongs to the log-GARCH class, the model is robust to spikes or jumps, allows for a rich variety of volatility dynamics without restrictive positivity constraints, can be estimated equation-by-equation by means of standard methods even in the presence of feedback, and allows for Dynamic Conditional Correlations (DCCs) that can - optionally - be estimated subsequent to the volatilities. We use the model to study the hourly day-ahead system prices at Nord Pool, and find extensive evidence of periodic volatility and volatility feedback. We also find that volatility is characterised by (positive) leverage in half of the hours, and that a DCC model provides a better fit of the conditional correlations than a Constant Conditional Correlation (CCC) model.

CO075 Room MAL 416 VOLATILITY MODELING AND DERIVATIVES PRICING Chair: Juan-Pablo Ortega

C1395: **Option pricing and hedging with one-step Kalman filtered factors in non-affine stochastic volatility models**
Presenter: **Juan-Pablo Ortega**, University St. Gallen, Switzerland
Co-authors: Alex Badescu, Lyudmila Grigoryeva
An innovative Kalman-based estimation technique is introduced for a non-affine auto-regressive stochastic factor model with non-predictable drift which allows to account for leverage effects. More specifically, we adopt a one-step unscented filtering approach which circumvents the use of the Kalman gain that produces a poor performance for this kind of models. New pricing and hedging strategies are proposed for contingent products that have this model for the underlying asset by introducing a volatility dependent exponential linear pricing kernel with stochastic risk aversion parameters. This technique proves to outperform standard GARCH and Heston-Nandi based strategies in terms of a variety of considered criteria in an empirical exercise using historical returns and options data.

C1362: **Two step modified-QML estimation for NIG-GARCH Processes**
Presenter: **Fanirisoa Zazaravaka R Hasinavonizaka**, Pantheon Sorbonne, France
Co-authors: Christophe Chorro
A two step Modified-Quasi Maximum likelihood (QML) procedure is proposed to estimate GARCH models with Normal Inverse Gaussian Distribution innovation (NIG). We provide also a comparison of GARCH-HN and GARCH-GJR models based on their capabilities of volatility modelling and forecasting abilities. Using our two step modified-QML estimation procedure, we model the volatility of S&P500 using option-returns and VIX-returns and compare the performance of our model with Gaussian innovations. Our results suggest that improvements of the overall estimation are achieved when Modified-QML Estimation are used with VIX-returns and when NIG distribution is taken into account in the conditional variance. Moreover, it is found that NIG-GARCH-GJR allows better forecasts than Gaussian-GARCH-HN, Gaussian-GARCH-GJR and the NIG-GARCH-HN. Finally, increased performance of the forecasts is clearly observed when using non-normal distributions and numerical studies confirm the advantages of the proposed approach.

C1801: **GAMM style volatility modeling**
Presenter: **Giacomo Bormetti**, University of Bologna, Italy
Co-authors: Giulia Livieri, Fulvio Corsi
In modeling daily volatility with high-frequency data, the econometrician faces two serious limitations: Realized measures are contaminated by noise and the overnight volatility is not observable. Both effects undermine forecasting ability and pricing performances. Inspired by the Generalized Autoregressive Method of Moments, we introduce a new reduced-form model for price returns which solves both issues. The observed close-to-open and open-to-close returns and realized volatilities step into the model via orthogonal conditions. As quantities with zero conditional

expectation, the latter drive the dynamics of the latent volatility components and jump intensity. Our observation-driven specification is tailored for an effective and computationally undemanding filtering, that reveals the wandering behaviour of the overnight component and a sizeable increase of persistence of the latent volatility. Remarkably, the model belongs to the class of affine processes and thus inherits all advantages deriving from analytical tractability. By means of an extensive analysis of S&P500 Futures time series and S&P500 Index options spanning more than two decades, we report the superior performances of our approach in comparison with competitor models.

C0697: **On the forecasting ability of option implied risk-neutral distributions**
Presenter: **Maria Magdalena Vich Llompart**, University of Balearic Islands, Spain
Co-authors: Antoni Vaello Sebastia
The forecasting ability of risk-neutral densities (RNDs) estimated using either parametric (mixture of two Log-Normal distributions) and non-parametric methods (kernel and splines) for different time horizons is analyzed. Traditional tests for the forecasting ability rely on restrictive assumptions (mainly normality and independence). In order to overcome these problems, we calculate block-bootstrap-based critical values. We consider RNDs on three US indexes, S&P500, Nasdaq 100 and Russell 2000 for a long series of data, ranging from 1996 to 2015, which is of special interest since it encompasses two major crisis. Differently to existent literature, our results conclude failure to reject their forecasting ability, being these results consistent across the different indexes and methodologies. We also analyze the fit of the tails of the RNDs separately, finding that, in general, they tend to overestimate the frequency of occurrence of events in the left tail (losses).

CO089 Room MAL 421 ADVANCES IN TIME SERIES MODELLING, FORECASTING, AND FORECAST EVALUATION Chair: Richard Luger

C1096: **Forecast evaluation and dynamic panels**
Presenter: **Charles Saunders**, University of Western Ontario, Canada
Calculation of forecast evaluation statistics for time series models requires a large number of predicted observations for appropriate inference when comparing models. Panel data can be limited to few time periods, so dynamic panel estimation methods correct for the well-known incidental parameter bias problem. A panel Diebold-Mariano statistic is constructed that includes the cross-sections of the panel. This allows for valid forecast inference with a minimum of a single cross-section of predicted errors. The panel Diebold-Mariano statistic is examined via simulation for critical values for both stationary and nonstationary models, finding that the standard normal critical values are appropriate for stationary panels with Large-N. The simulation study examines the effect of using a consistent (moment-based) or an inconsistent (fixed-effects) estimator. Forecast evaluation of nonstationary panels lead to size distortions for some GMM estimators, that are known to perform poorly at the unit boundary.

C0695: **Comparing out-of-sample forecasts against a random walk: Exact tests with application to exchange rates**
Presenter: **Sermin Gungor**, University of Western Ontario, Canada
Co-authors: Richard Luger
An exact inference procedure is developed to test the null hypothesis that a given model's out-of-sample forecasts are no better than a random walk. The proposed Monte Carlo resampling-based procedure accounts for parameter uncertainty and leaves unrestricted the estimation scheme, the forecast horizon, and the forecast evaluation criterion. In contrast to a previously developed Monte Carlo forecast evaluation procedure, our approach is distribution-free and free of nuisance parameters. A simulation study demonstrates the fact that the proposed procedure achieves size control and has good power in comparison to alternative approaches, including various bootstrap methods. We apply the new procedure to test the out-of-sample predictive ability of economic fundamentals in empirical exchange rate models.

C0657: **Modelling higher moments and density forecasting: A comprehensive look**
Presenter: **Xiaochun Liu**, University of Alabama, United States
Co-authors: Richard Luger
Many GARCH-type models have been proposed in the literature for the higher moments of financial returns and their conditional distributions. We examine comprehensively whether these models yield better out-of-sample density forecasts. Among a wide range of specifications for autoregressive conditional volatility, skewness, and kurtosis, we find that the most promising approach rests on a decomposition of returns into their signs and absolute values. This approach specifies the joint distribution of the return components by combining a dynamic binary choice model for the signs, a multiplicative error model for the absolute values, and a dynamic copula function for their interaction. This flexible specification captures well the time-varying conditional skewness process and provides more accurate density forecasts than competing models, especially for the left tail of financial returns.

C1098: **Dynamic interaction between sovereign credit rating events and credit default swaps**
Presenter: **Richard Luger**, Laval University, Canada
An empirical framework is proposed for modelling and assessing the dynamic interaction between credit rating events (changes in the rating or outlook) and CDS spreads, with a particular focus on the sovereign debt market. The approach combines: (i) a GARCH model for CDS spread changes; (ii) a dynamic logit model for credit rating events; and (iii) a dynamic copula model for the non-linear dependence structure between CDS spreads and credit rating events. The end result is a complete characterization of the dynamic dependencies, spillovers, and feedback effects between sovereign credit rating events and CDS spreads. We apply the new methodology to five Euro area countries (Greece, Ireland, Italy, Portugal, and Spain) after the collapse of Lehman Brothers and find evidence of asymmetric propagation effects of credit rating announcements (by Standard & Poor's, Moody's, and Fitch) for one country on the sovereign CDS spreads and credit ratings of other countries.

CO110 Room MAL 532 TOPICS IN FINANCIAL ECONOMETRICS Chair: Joern Sass

C0666: **A performance comparison of long-term (optimal) investment strategies based on improved smoothing techniques**
Presenter: **Michael Scholz**, University of Graz, Austria
Co-authors: Jens Perch Nielsen, Stefan Sperlich, Enno Mammen
In long-term investment products it is important to understand the underlying financial risk of the optimal investment profile. Various performance measures, for example, the Sharpe-ratio were proposed to evaluate those investment strategies in practice. We provide an improved estimator for the Sharpe-ratio which includes prior knowledge in the estimation process of conditional mean and variance function in a predictive regression model focusing on nonlinear relationships between a set of covariates. In an applied part, we compare different investment strategies (e.g. simple, static optimal, or dynamic optimal) based on our improved estimators using annual data of the S&P500 in a period from 1872 to 2015.

C0686: **Dynamic modeling of measures of credit quality**
Presenter: **Laura Vana**, WU Wirtschaftsuniversitaet Wien, Austria
Credit risk modeling including the measurement of credit quality has been intensively investigated by academics and practitioners over the past decades. The aim is to contribute to this field of research by developing a framework for jointly modeling firm failures (e.g., bankruptcies) and ordinal credit ratings as outcomes. This model, unlike prior work, simultaneously incorporates failures and credit ratings and allows inference about the quantitative relationships between these outcomes by simultaneously making use of both sources of information. A dynamic latent trait

model is employed for the latent creditworthiness process underlying the rating and failure observations. Time dependent processes are defined for the systematic and idiosyncratic development of credit quality. Failure, firm-level and stock price data for publicly traded North American companies as well as issuer credit ratings from the big three rating agencies (S&P, Moody's and Fitch) are collected and analyzed to illustrate the proposed framework. Full Bayesian inference is perfomed using MCMC methods.

C0694: **Adjustable network reconstruction with applications to CDS exposures**
Presenter: **Luitgard Veraart**, London School of Economics, United Kingdom
Co-authors: Axel Gandy

The problem of reconstructing weighted directed networks from the total in- and out-weight of each node is considered. This problem arises, for example, in the analysis of systemic risk of partially observed financial networks. Typically, a wide range of networks is consistent with this partial information. We develop an empirical Bayesian methodology that can be adjusted such that the resulting networks are consistent with the observations and satisfy certain desired global topological properties such as a given mean density. Furthermore, we propose a new fitness based model within this framework. We apply our methodology to a novel data set containing 89 financial networks of credit default swap exposures. The performance of the reconstruction methodology is very good under a wide range of performance criteria and also compared to other existing reconstruction methods. In particular, we are able to reconstruct the degree distribution of the underlying networks with remarkable precision if a good estimate of the true density of the underlying network is available.

C0824: **Large scale portfolios under transaction costs and model uncertainty: Mixing of high and low frequency information**
Presenter: **Stefan Voigt**, WU (Vienna University of Economics and Business), Austria
Co-authors: Nikolaus Hautsch

A Bayesian sequential learning framework is proposed for high-dimensional asset allocations under model ambiguity and parameter uncertainty. We consider portfolio allocations maximizing predictive expected utility after transaction costs, optimally balancing implementation shortfall and adjustments due to updated information. The unifying framework allows for time-varying mixtures of predictive return distributions which may exhibit fat tails, resulting from high- and low-frequency data. The model is estimated via MCMC methods and allows for a wide range of data sources as inputs. We consider predictive models resulting from high-dimensional Wishart approaches for high-frequency based blocked realized kernels, low-frequency based multivariate stochastic volatility factor models and regularized daily covariance estimates. Employing the proposed framework on a large set of NASDAQ-listed stocks,we observe that time-varying mixtures of high- and low-frequency based return predictions significantly improve the out-of-sample portfolio performance compared to individual models and outperform the naive 1/N-allocation in terms of Sharpe ratio and utility-based measures. Bootstrapping the optimization procedure shows that our results are robust with respect to the choice of the asset universe. We show that regularization of turnover is crucial in large dimensions and illustrate that the relative contribution of high-frequency data and low-frequency data strongly varies over time.

CO502 Room MAL 538 NOWCASTING METHODS IN MACROECONOMETRICS Chair: Clement Marsilli

C0458: **Markov-Switching three-pass regression filter**
Presenter: **Danilo Leiva-Leon**, Banco de España, Spain

A new approach is introduced for the estimation of high-dimensional factor models with regime-switching factor loadings by extending the linear three-pass regression filter to settings where parameters can vary according to Markov processes. The new method, denoted as Markov-Switching three-pass regression filter (MS-3PRF), is suitable for datasets with large cross-sectional dimensions since estimation and inference are straightforward, as opposed to existing regime-switching factor models, where computational complexity limits applicability to few variables. In a Monte-Carlo experiment, we study the finite sample properties of the MS-3PRF and find that it performs favorably compared with alternative modelling approaches whenever there is structural instability in factor loadings. As empirical applications, we consider forecasting economic activity and bilateral exchange rates, finding that the MS-3PRF approach is competitive in both cases.

C1041: **Nowcasting Indian GDP**
Presenter: **Daniela Bragoli**, Universita Cattolica Milano, Italy
Co-authors: Jack Fosten

A nowcasting model is proposed for the Indian real GDP growth rate which uses the flow of relevant information to update predictions on a daily basis. Based on a dynamic factor model, our method can serve as a timely barometer to track the Indian development process. There are several challenges faced when nowcasting GDP in developing economies such as India. The first challenge is to proxy important missing variables such as international trade in the service sector. Our novel solution augments a baseline model with series on US and Euro-area output which improves predictions, particularly during the 2008-2009 global crisis. The second challenge is the impact of sizeable revisions to the GDP data. We construct a new series for real-time Indian GDP using press releases from the Central Statistics Office (CSO), finding that data revisions have a non-trivial influence on our results. Therefore, caution should be taken when evaluating predictions using the preliminary GDP release.

C1181: **Forecasting with Bayesian adaptive penalized mixed-frequency regressions**
Presenter: **Clement Marsilli**, Banque de France, France
Co-authors: Matteo Mogliani

A new approach is proposed to forecast with mixed-frequency regressions (MIDAS) that address the issues of estimation and variable selection in presence of a large number of predictors. Our approach is based on adaptive penalized regression models (Lasso, Group Lasso, and Elastic-Net) and relies on Bayesian techniques for estimation. In particular, the penalty parameters driving the model shrinkage are automatically fine-tuned via an adaptive MCMC algorithm, which is computationally efficient compared to the standard EM algorithm. Simulations show that the proposed penalized MIDAS models significantly outperform a benchmark represented by an optimal combination of single-predictor MIDAS regressions. When applied to US GDP, the results suggest that our models produce significant out-of-sample predictive gains compared to several alternative models.

C1622: **Improving the classification accuracy of the logit model: The case of US bank failures**
Presenter: **Alexander Kostrov**, University of St. Gallen, Switzerland
Co-authors: Francesco Audrino, Juan-Pablo Ortega

A MIDAS-style weighting scheme is introduced for constructing flow predictors in financial studies. MIDAS is introduced in the context of the logit model and we then address the issue of classification accuracy evaluation in the presence of severe class imbalance in the data. A 'risk group' approach is presented as a better alternative compared to the standard indicators of the classification accuracy, which are misleading for imbalanced data. Re-weighting of observations in the log-likelihood function is suggested to mitigate the class-imbalance problem, where the cross-validation selects an optimal weight for the rare class. In the empirical part of the study, we apply these innovations to improve a well-established logit model for predicting US bank failures in 2004-2016. The gain in forecasting accuracy is confirmed to be both statistically and economically significant. In our setting, MIDAS-style weights characterize the relationship between the probability of individual US bank failures and past values of an explanatory flow variable.

CO314 Room MAL 539 ENERGY MARKETS Chair: Marco Lorusso

C0683: **Leverage and the oil industry: Analysis on the firm and production level**
Presenter: **Johannes Lips**, Justus-Liebig University Giessen, Germany
The purpose is to analyze the relationship between debt and the production decision of companies active in the exploration and production of oil and gas in the United States. Over the last couple of years, the development and application of innovative extraction methods, like hydraulic fracturing and horizontal drilling, led to a considerable increase in US oil production. In connection with these technological changes, another important economic development in the oil industry has been identified: largely debt-driven investments in the oil sector. Additionally, the rising prices in the commodities markets until mid 2014 led to higher asset valuation and thus to higher return expectations fueling a virtuous circle and increasing the oil and gas production, especially in the US. The first research question is whether debt and leverage affects production decisions of oil companies. The second research question then is, if the technological changes in the industry and the increased indebtedness of US oil companies led to a markedly different reaction in their production decision between the episodes of oil price decline in 2008 and 2014. A potential reason for the absence or delay in cutting back production after the price drop in 2014 could be supposedly higher leverage prior to the price decline. These questions are econometrically addressed using a novel dataset combining financial data on publicly listed firms and their production data on well level.

C0907: **On energy labels for dwellings and retrofitting: Some evidence from the English energy performance certificate**
Presenter: **Mirko Moro**, University of Stirling, United Kingdom
How to induce households to install energy efficient technology remains a puzzle. Could an energy labelling requirement for residential real estate help? We propose that the salient color-letter grades on the English Energy Performance Certificate (EPC) served as targets, motivating vendors to invest in energy efficiency. In the post-requirement years we find a cluster of homes with energy efficiency scores just above the D-grade threshold. This cluster was not present prior to the requirement, replicates in an independently-drawn representative sample, is largest for homes that had been on the market since implementation of the requirement. We conclude that the EPC requirement induced investment, and hence that energy efficiency labels have potential to green the housing stock. We infer from our analysis how the design of the EPC could be altered to motivate greater investment in energy efficiency.

C0999: **Oil prices and informational frictions: The time-varying impact of fundamentals and expectations**
Presenter: **Bing Xu**, Heriot-Watt University, United Kingdom
Co-authors: Joseph Byrne, Marco Lorusso
Informational frictions are considered when modelling the time-varying relationship between crude oil prices, traditional fundamentals and expectations. Informational frictions force a wedge between oil prices and supply and/or demand shocks, especially during periods of elevated risk aversion and uncertainty. In such a context expectations can be a key driver of oil price movements. We utilize a variety of proxies for forward-looking expectations, including business confidence, consumer confidence and leading indicators. In addition, we implement a time-varying parameter approach to account empirically for time-varying informational frictions. The results illustrate firstly that oil supply shocks played an important role in both the 1970s and coinciding with the recent shale oil boom. Secondly, demand had a positive impact upon oil prices, especially from the mid-2000s. Finally, we provide evidence that oil prices respond strongly to expectations but the source of the shock matter: business leaders expectations are positively related, while markets expectations are not strongly linked to oil prices.

C0966: **Oil and fiscal policy: Panel regime-switching country analysis**
Presenter: **Marco Lorusso**, Heriot-Watt University, United Kingdom
Co-authors: Francesco Ravazzolo, Roberto Casarin, Hilde Bjornland
The focus is on the response of government policy to oil price changes in a set of oil exporter countries. Our country sample consists of both advanced and developing economies. The main research questions are: (a) do oil price changes affect fiscal regimes of oil exporter countries? (b) what is the relationship between expensive/contractionary fiscal policies and oil price fluctuations? (c) do world oil production and global economic activity influence government policies of oil exporter countries? (d) what are the links between government spending, tax revenues, public and private employment and oil prices in these countries? As empirical framework we use a Bayesian panel model VAR for mixed frequency data whose parameters can change over time according to a Markov process. We believe that this approach presents several advantages. Firstly, we identify regimes depending on the property of the estimated parameters and avoid critiques about the identification of standard VARs with indirect methods such as a recursive scheme. Secondly, our empirical framework allows us to distinguish between countries adopting oil fiscal rules in order to shelter their economies and countries that increase their spending when their revenues are high. Finally, the flexibility of our model gives us the possibility to compare fiscal policies of a relative large set of countries that would not be possible adopting a standard VAR model.

CO242 Room MAL 540 HIGH FREQUENCY FINANCIAL MODELLING Chair: Vitali Alexeev

C1045: **High-frequency quoting and liquidity commonality**
Presenter: **Riccardo Borghi**, Cass Business School, United Kingdom
The last years have seen the growth of the high-frequency trader (HFT): a firm that uses proprietary capital to act as market maker on multiple securities engaging in high-frequency quoting (HFQ). The relationship between HFQ and intraday liquidity commonality, i.e. the explanatory power of common liquidity factors on stock liquidity, is investigated to test if HFQ activity increases market interconnectedness. The sample comprises all trades and best quotes' updates for the FTSE100 stocks from January 2010 to December 2014, traded on the London Stock Exchange. An upgrade of the trading systems of the London Stock Exchange on February 2011 is used to identify an exogenous positive shock to HFQ. The empirical results suggest that HFTs are most active before 12pm when stocks are less liquid and market makers are needed the most, while HFQ decreases steadily after 12pm and it is the lowest at the end of the day. Furthermore, liquidity is the highest and the most common at the end of the day, which might suggest the presence of informed market makers. So far, no change in the intraday relationship between HFQ and liquidity commonality has been found after the update of the LSE trading system.

C0357: **Quantile regression models with factor-augmented predictors and time-varying factor loadings**
Presenter: **Alev Atak**, City University London, United Kingdom
Co-authors: Yiguo Sun, Yonghui Zhang
A semiparametric quantile regression model with factor-augmented predictors and time-varying factor loadings is developed, where the time-varying factor loadings is allowed to change across quantile regressions at different probability masses while taking the latent factors fixed. We propose a two-stage procedure. In the first step, we simultaneously estimate the latent factors and time-varying factor loadings, defined by a nonparametric smooth function, using a local version of the principal component method. In the second step, we develop our quantile regression model with factor-augmented predictors that is derived in the first step. The proposed method extracts and combines distributional information across different semiparametric quantile regression models. Results of Monte Carlo simulations demonstrate that the proposed criterion performs well in a wide range of situations. We also apply our model to investigate a U.S. macroeconomic data set and find strong evidence of heterogeneity in dynamic responses.

C0447: **Daily vs intraday risk assessment using asynchronous tick-by-tick data**
Presenter: **Simona Boffelli**, Bergamo University, Italy
Co-authors: Giovanni Urga
When tick-by-tick data are available, risk evaluation conducted at high-frequency is shown to outperform their counterparts executed with the traditional daily frequency. To this purpose, it is crucial to adopt an appropriate synchronization scheme and a correct estimator of the integrated covariance matrix when working in a multivariate framework. We evaluate a portfolio of European government bonds over the period 1st June 2007 31st May 2012 focusing on a trading strategy based on risk control measure and a risk management backtesting exercise using unilevel, multilevel Value-at-Risk tests and Tail Risk measures. We provide clear cut evidence of the benefit of exploiting the additional information in the intraday data with respect to daily frequency. Both trading strategy and risk management backtesting tests more accurately forecast VaR when high-frequency data are used.

C0327: **Warping time: Improving efficiency of tick-by-tick data in portfolio optimisation**
Presenter: **Vitali Alexeev**, University of Technology Sydney, Australia
Co-authors: Giovanni Urga
A unified framework is developed that allows analysis of unevenly spaced tick-by-tick data. Up to 90 percent of tick-by-tick data is lost during pre-processing of the data to fit existing models in finance applications. Compared to existing methods, the proposed approach avoids unnecessary loss of observations and allow for flexible time shifts by warping the time domain. The resulting framework is capable of direct analysis of tick-by-tick financial data while simultaneously addressing the main empirical issues identified in the contemporary high-frequency literature for such data (asynchronous trading and microstructure noise). Correlation structures estimated based on properly aligned time series at high frequencies allow for improved portfolio allocation strategies and decision-making process for investment professionals.

CO176 Room MAL 541 FUNDS PERFORMANCE MEASUREMENT Chair: Spyros Vrontos

C1573: **Unveiling the risk profile of funds of hedge funds**
Presenter: **Christos Argyropoulos**, Lancaster University, United Kingdom
Co-authors: Christos Argyropoulos, Ekaterini Panopoulou, Spyros Vrontos
Hedge funds implement complex dynamic investment strategies with option like non-linear payoffs, involving leverage and short-selling in pursue of absolute returns. Fund of funds, as portfolios of hedge funds, offer the same investment characteristics and at the same time diversify across individual hedge funds. In order to formulate portfolios of hedge funds, fund of funds managers perform their due diligence without disclosing their selection criteria, the portfolios composition and their associated risk. We compare the risk return profile of hedge funds and fund of funds in order to evaluate the added value of fund of funds in relation to the underlying hedge funds. We construct decide portfolios of hedge funds and fund of funds in order to create benchmarks of the risk return profile of both investment vehicles. In addition, we propose optimal fund of funds strategies in order to create a portfolio of hedge funds that minimizes downside risk. Our findings suggest that for the low levels of risk, hedge funds are less risky than fund of funds while they provide better average returns. On the other hand, for high levels of risk the fund of funds provide a diversification effect at the cost of significantly reduced returns. More importantly, our proposed fund of funds strategy dominates the corresponding risk returns profile of funds of funds and individual hedge funds.

C1667: **Funds performance evaluation: A review and a comparison**
Presenter: **Spyros Vrontos**, University of Essex, United Kingdom
Extending previous work on mutual fund pricing, the idea of modelling the conditional distribution of mutual fund returns using a number of alternative models is introduced, taking into account the stylised facts of mutual fund return series. We evaluate mutual fund performance using multifactor asset pricing models, with the relevant risk factors being identified through standard model selection techniques. We explore potential impacts of our approach by analyzing individual mutual funds and show that it can be economically important.

C1483: **Examination of hedge fund performance persistence over long-term period using a non-parametric approach**
Presenter: **Ekaterina Ipatova**, Roehampton Bisiness School, United Kingdom
Co-authors: Kaizad Doctor
Persistence in hedge fund performance is hotly contested by the academic community. Evaluating persistence in fund performance is generally carried out using parametric and non-parametric methods. Procedures that identify superior fund performance after mitigating the possibility of lucky fund managers are used in identifying the set of funds that are generally available in the literature. The methodologies have varying advantages with the False Discovery Rate (FDR) method being less conservative than the Family Wise Error Rate (FWER) however parametric methods are also added to the mix. A comparison of the methodologies has not been carried out in literature. Varying horizons, mitigating survivorship bias and out-of-sample analysis have all been incorporated in my study. A variety of portfolios were used to evaluate the FOF with traditional performance evaluation methods ranging from non-parametric Sharpe ratios and Drawdowns. We find that funds constructed using non-parametric methods outperform those using parametric methods and in general exhibit superior outperformance when compared to conventional portfolios.

CG123 Room MAL 151 CONTRIBUTIONS IN GARCH Chair: Alan Hawkes

C0239: **A GARCH-Hawkes jump model: Self-excitation and calibration**
Presenter: **Alan Hawkes**, Swansea University, United Kingdom
Co-authors: Jing Chen, Steve Yang
With consideration of increased observations on contagion effects of market events occurring in financial markets, it is argued that classic diffusion models and/or their extension incorporating Poisson or Levy jumps are not sufficient to best describe the non-linearity of financial time series. Instead, we propose to incorporate an intensity based model and the simplest but most effective choice is a one-dimension Hawkes process self-excitation. Our aim is to establish such a model with practicality that leads us to focus on calibrating the Hawkes jump model and comparing its effectiveness with the NGARCH model, experimenting over a long period of intraday price series of S&P 500. During the model estimation, we use a Monte Carlo EM algorithm to achieve the optimisation and overcome the issue of obtaining the marginal distribution of unobserved data. We also use a simulated annealing optimisation algorithm to further enhance the optimisation process. Finally, we compare the Heston and Hawkes models to validate our proposal through better achieved model parameters without digging into complicated option pricing or establishing volatility forecasting.

C0386: **Simultaneous estimations of the parameters regression with Realized-GARCH errors**
Presenter: **Hisseine Saad Mahamat**, University of Montpellier, France
Co-authors: Roman Mestre, Michel Terraza
The systematic risk of an equity (measured with the Beta) is estimated by the market line equation. According to the OLS hypothesis, the estimation is robust and residuals are normally white-noise process. However, various papers show that the BETA estimator cannot be BLUE because there are many statistical anomalies in the residuals (heteroskedasticity, autocorrelation and non-normality) proving then that the the model does not

hold. The Beta value can be different and the estimation more robust if we re-estimate it by taking into consideration the heteroskedastic nature of the residuals. We base our calculation on the Societe Generale risk premium (a French Bank Equity) and CAC40 premium (French Stockmarket Index) for the daily period from 2005-2015. We use a new GARCH process; called Realized-GARCH, to model the errors variance and we include it in the Market line estimation. The asymmetric nature of the GARCH is also tested; in order to compare the result with other GARCH process such as EGARCH, GJR-GARCH. To select the best GARCH process to correct the Beta estimation, we compare the result about the Beta value and the residuals characteristics. We find that the Beta value with GARCH errors is different and the residuals characteristics are better.

C1728: **Minimum variance hedging when using implied covariances**
Presenter: **Andres Algaba**, Vrije Universiteit Brussel, Belgium
Co-authors: Kris Boudt, Steven Vanduffel
The minimum variance hedge ratio is defined as the ratio of the conditional covariance between spot and futures returns to the conditional variance of futures returns. We recommend to imply the conditional covariance from the conditional variance of the weighted sum of spot and futures returns to avoid the modelling restrictions and computational challenges in estimating traditional BGARCH models. An optimal weight can be derived by minimizing the scaled variance of the sample variance estimator of this weighted sum. In a simulated and empirical application, we find that our approach performs as well as BGARCH models.

C1791: **Volatility transmission in multiply overlapping trading zones**
Presenter: **Andreas Masuhr**, University of Munster, Germany
Previous volatility spillover models use artificially non overlapping trading zones to identify sources of volatility transmission between these zones. The problem of non overlapping zones is overcome using a copula GARCH approach that allows for multiple overlaps between zones incorporating vine copulas to flexibly model the dependence structure and to meet stylized facts of return data. To handle the relatively large parameter space, the model is estimated by Bayesian methods using a differential evolution MCMC approach. Results are presented for an exchange rate data set containing three trading zones with a single overlap, as well as for simulated data for a multiple overlap setting. Finally, the hypothesis that volatility transmits from preceding trading days as well as from preceding trading zones is supported.

CG008 Room MAL 152 CONTRIBUTIONS IN NONLINEAR TIME SERIES Chair: Cristina Amado

C1513: **A score-driven smoother for general state-space models**
Presenter: **Giuseppe Buccheri**, Scuola Normale Superiore, Italy
Co-authors: Giacomo Bormetti, Fulvio Corsi, Fabrizio Lillo
A simple approximate smoother is introduced which extends the score-driven estimation approach to also include present and future observations. The newly proposed Score-Driven Smoother (SDS) can be used to improve the estimation of time-varying parameters in nonlinear non-Gaussian state-space models. In contrast to complex and computationally demanding simulation-based methods, the SDS has the same simple structure of the Kalman backward smoothing recursion but uses the score of the true observation density. Through an extensive Monte Carlo study, we provide evidence that the performance of the approximation is very close (with average differences of 2% in mean square errors) to that of simulation-based techniques and is superior to other approximate methods. Empirically, the effectiveness of the SDS is shown in recovering accurate estimates of time-varying volatilities and correlations of inflation rates in the euro area.

C1582: **On partial-sum processes of ARMAX residuals**
Presenter: **Steffen Groenneberg**, BI Norwegian Business school, Norway
Co-authors: Benjamin Holcblat
General and versatile results are established regarding the limit behavior of the partial-sum process of ARMAX residuals. Illustrations include ARMA with seasonal dummies, misspecified ARMAX models with autocorrelated errors, nonlinear ARMAX models, ARMA with a structural break, a wide range of ARMAX models with infinite-variance errors, weak GARCH models and the consistency of kernel estimation of the density of ARMAX errors. Our results identify the limit distributions, and provide a general algorithm to obtain pivot statistics for CUSUM tests.

C1626: **Time series analysis using the radial basis functions: Application to the US economy**
Presenter: **Nobuyuki Kanazawa**, Hitotsubashi University, Japan
A flexible nonlinear method is proposed for studying the time series properties of macroeconomic variables. We focus on a class of Artificial Neural Networks (ANN) called the Radial Basis Functions (RBF), which is capable of producing history- and shock-dependent impulse responses without imposing a strong functional form assumption. To assess the validity of the RBF approach in a macroeconomics time series analysis, we conduct a series of Monte Carlo experiments using a data generated from a nonlinear New-Keynesian (NK) DSGE model. We find that the RBF time series can uncover the nonlinear NK economic structure from the simulated data whose length is as small as 300 (quarters). Finally, we apply this RBF time series method to the US macroeconomic data from 1948-2015 and show that the response of real GDP growth to utilization-adjusted TFP shocks is negative after 2008 when the nominal interest rates hit zero. The finding is consistent with the prediction of the New Keynesian model.

C0347: **New methods for model diagnostics in multiplicative error models**
Presenter: **Indeewara Perera**, Monash University, Australia
Co-authors: Mervyn Silvapulle
The recent literature on financial time series analysis has devoted considerable attention to non-negative random variables, such as the duration between trades at a stock exchange, realized volatility, volume transactions and squared or absolute returns. The class of models, referred to as the multiplicative error models, is particularly suited to model such non-negative time series. A multiplicative error model (mem) decomposes the non-negative time-series variable of interest into the product of its conditional mean and a multiplicative error term. Kolmogorov-Smirnov type tests are developed for testing the parametric specification of a given mem which consists of separate parametric models for the conditional mean and the error distribution. The limiting distributions of the test statistics are model-dependent and not free of nuisance parameters, and hence critical values cannot be tabulated for general use. A bootstrap method is proposed for computing the p-values of the tests, and is shown to be consistent. To this end, a new general result on asymptotic uniform expansions is established for a class of randomly weighted residual empirical processes. This is a useful result in its own right. The proposed tests are shown to have nontrivial asymptotic power against a class of root-n local alternatives. The tests performed well in a simulation study, and are illustrated using a data example on realized volatility.

CG255 Room MAL 153 CONTRIBUTIONS IN QUANTITATIVE INVESTING Chair: Romeo Tedongap

C0209: **Quantitative investing a stock selection system for Europe**
Presenter: **Ramon Bermejo Climent**, Universidad Pontificia Comillas (Madrid, Spain), Spain

The universe of European corporate data for the 1990-2017 period are used to demonstrate that equity portfolios built following value, profitability and momentum factors achieve higher returns that their market benchmark both in absolute and risk-adjusted terms. These three factors have been analysed in detail mainly for the US market while we focus on testing their persistency, pervasion and robustness across the European equity markets using an exclusive data set. Following this test, we are able to construct a quantitative systematic investment strategy that selects the stocks in the portfolios through an algorithm based on value, profitability and momentum. We demonstrate that our portfolios, rebalanced on a yearly basis, overperform the European equity benchmark over the long term, especially when the three factors are combined together.

C1326: **On the stock analysts' capacity to predict volatility**
Presenter: **Ivan Medovikov**, Brock University, Canada

Investment value of sell-side analyst recommendations from the standpoint of portfolio risk is assessed. We match I/B/E/S consensus recommendations issued for U.S.-listed equities during January 2015 with realized volatility of daily returns up to one year following recommendation issue. Using a flexible semiparametric copula model we find recommendation levels to be associated with subsequent changes in volatility, suggesting that analyst ratings can help manage portfolio risk. This relationship appears to be asymmetric and is most pronounced among the best-rated securities which experience largest volatility declines after recommendation issue. These effects are conditional on recommendation changes.

C0873: **Supercointegrated**
Presenter: **Antoni Vaello Sebastia**, University of Balearic Islands, Spain
Co-authors: Isabel Figuerola-Ferretti, Pedro Serrano, Tao Tang

The performance of pair-trading portfolios is examined when sorted by the level of cointegration of their constituents. The supercointegrated portfolio, that formed by pairs at 1% confidence level of cointegration tests, exhibits a superior out-of-sample performance than simple buy-and-hold and passive investments. The time-varying risk of the pairs strategy is linked to aggregate market volatility. A positive risk-return relationship of the strategy is also found.

C1811: **Linguistic views in portfolio selection: Black-Litterman model extension**
Presenter: **Aleksandra Rutkowska**, Poznan University of Economics and Business, Poland
Co-authors: Marcin Bartkowiak

The Black-Litterman model provides a framework in which two sources of information about the expected returns can be combined. The first source of information is the expected returns that follow from the Capital Asset Pricing Model and the second source of information are the views held by the investors. Since the investors views about future asset return are always subjective, imprecise and often express in natural language, we propose extension includes two elements important from the point of view of application: linguistic information and the multiple experts views. This is possible owing to use ill-known random variables. We introduce the model extension step-by-step and present empirical example.

Sunday 17.12.2017 10:50 - 12:55 Parallel Session H – CFE-CMStatistics

EO613 Room CLO B01 SCALABLE BAYESIAN METHODS FOR LARGE DATA PROBLEMS Chair: Mattias Villani

E0931: **Exact subsampling in MCMC using piecewise deterministic Markov processes**
Presenter: **Joris Bierkens**, Delft Institute of Applied Mathematics, Netherlands

Markov chain Monte Carlo methods provide an essential tool in statistics for sampling from complex probability distributions. While the standard approach to MCMC involves constructing discrete-time reversible Markov chains whose transition kernel is obtained via e.g. the Metropolis-Hastings algorithm, there has been recent interest in alternative schemes based on piecewise deterministic Markov processes (PDMPs). One such approach is based on the Zig-Zag process, which proved to provide a highly scalable sampling scheme for sampling in the big data regime, as it allows for subsampling without modifying the posterior distribution. We will present a broad overview of these methods along with some theoretical results.

E1123: **Exact subsampling MCMC**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: Matias Quiroz, Mattias Villani, Minh-Ngoc Tran

Speeding up Markov Chain Monte Carlo (MCMC) for data sets with many observations by data subsampling has recently received considerable attention in the literature. Most of the proposed methods are approximate, and the only exact solution has been documented to be highly inefficient. We propose a simulation consistent subsampling method for estimating expectations of any function of the parameters using a combination of MCMC with data subsampling and a previous importance sampling correction for occasionally negative likelihood estimates. Our algorithm is based on first obtaining an unbiased but not necessarily positive estimate of the likelihood. The estimator uses a soft lower bound such that the likelihood estimate is positive with a high probability, and computationally cheap control variables to lower variability. Second, we carry out a correlated pseudo marginal MCMC on the absolute value of the likelihood estimate. Third, the sign of the likelihood is corrected using an importance sampling step that has low variance by construction. We illustrate the usefulness of the method with two examples.

E1271: **Efficient particle filter methods for diffusions**
Presenter: **Michael Pitt**, Kings College London, United Kingdom

Recent developments in Particle Metropolis Hastings methods applied to diffusions are investigated. The scaling issue for large time series is considered. The methods are illustrated on standard models and for large T diffusion models which typically arise in finance. The methodology is general and straightforwardly applied.

E0937: **Fast and scalable Bayesian spatial 3D priors for brain imaging**
Presenter: **Per Siden**, Linkoping University, Sweden
Co-authors: Mattias Villani

Gaussian Markov random field (GMRF) priors will be considered, which have been successfully used in many large scale spatial problems, thanks to the sparsity induced in the precision matrices. However, most commonly used inference methods rely on the sparse Cholesky factorization, which is not feasible for problems of very large size, as those arising in spatial whole-brain modeling of task-related fMRI data, normally containing hundreds of thousands of data points and parameters. We instead develop fast and scalable inference algorithms utilizing among other preconditioned conjugate gradient methods, which are used both for sampling (MCMC) and stochastic optimization (VB and MAP). The methods are applied to fMRI data using a model which also has a non-trivial temporal component, and show to be both faster and more accurate than previous methods.

E1814: **Bayesian generative adversarial networks**
Presenter: **Andrew Wilson**, Cornell University, United States

Through an adversarial game, the recently proposed generative adversarial networks (GANs) can implicitly learn rich distributions over images, audio, and data which are hard to model with an explicit likelihood. We will present a practical Bayesian formulation for unsupervised and semi-supervised learning with GANs. Within this framework, we use stochastic gradient Hamiltonian Monte Carlo for marginalizing parameters. The resulting approach can automatically discover complementary and interpretable generative hypotheses for data. Moreover, by exploring an expressive posterior over these hypotheses, we show that it is possible to achieve state-of-the-art quantitative results on major image classification benchmarks even with less than 1% of the labelled training data.

EO352 Room MAL B18 CAUSAL INFERENCE IN THEORY AND PRACTICE II Chair: Jonas Peters

E0271: **Counterfactual fairness**
Presenter: **Ricardo Silva**, UCL / The Alan Turing Institute, United Kingdom
Co-authors: Matt Kusner, Joshua Loftus, Chris Russell

Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending, hiring, and predictive policing. In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for example those of a particular race, gender, or sexual orientation. Since this past data may be biased, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices. We develop a framework for modeling fairness using tools from causal inference. Our definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if it the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of success in law school.

E0288: **Large-scale causal discovery from nonlinear time series datasets**
Presenter: **Jakob Runge**, Imperial College London, United Kingdom

Detecting causal associations in observational time series datasets is a key challenge for novel insights into complex dynamical systems such as the Earth system. Dependencies in such a high-dimensional dynamical system may involve time-delays, nonlinearity, and strong autocorrelations, which present major challenges for causal discovery techniques. Here we are interested in time-lagged causal discovery using conditional independence testing and address two major problems: (1) Low power due to high-dimensionality and (2) detection 'biases'. The latter refers to the problem that the detection power for individual links may depend not only on their causal strength, but also on autocorrelation and other dependencies. We introduce a method for large-scale, linear and nonlinear, time-delayed causal discovery. In extensive numerical experiments we find that our method yields more power than common methods and largely overcomes detection biases allowing to more accurately rank associations in large-scale analyses by their causal strength. We demonstrate the method on a large-scale climate time series dataset.

E0921: **Inference of instantaneous causal relations in multivariate linear time series by stabilizing conditional distributions**
Presenter: **Niklas Pfister**, ETH Zurich, Switzerland
Co-authors: Jonas Peters, Peter Buehlmann
The problem of inferring the causal variables of a response Y from a set of d predictors $(X^1,\dots,X^d)$ is investigated. Given that $(Y_t,X_t^1,\dots,X_t^d)$ is a multivariate linear time series, we want to identify causal effects. This includes time instantaneous and lagged effects and therefore extends Granger causality. In contrast to only considering lagged effects, inferring also the instantaneous effects is a much harder task as the causal direction is unknown a priori. We present a method that makes use of heterogeneity patterns (or non-stationarity) in the data to detect the instantaneous causal relations and show that it satisfies some desirable statistical properties. To illustrate practical applicability we apply our method to a data set related to the monetary policy of the Swiss National Bank.

E1813: **Low priced lunch in conditional independence testing**
Presenter: **Jonas Peters**, University of Copenhagen, Denmark
Co-authors: Rajen Shah
Testing for conditional independence between continuous random variables is considered to be a hard statistical problem. We provide a formalization of this result and show that a test with correct size does not have power against any alternative. It is thus impossible to obtain any guarantee if the relationships between the random variables can be arbitrarily complex. We propose a practical test that achieves the correct size if the conditional expectations are smooth enough such that they can be estimated from data.

E1704: **Binary exposure and longitudinal cognition outcomes in the presence of non-ingorable dropout and death**
Presenter: **Maria Josefsson**, Centre for Demographic and Ageing Research, Sweden
Co-authors: Michael Daniels
G-computation, g-estimation and inverse probability weighting, have been proposed as alternatives to regression for causal inference of time-varying exposures. Although missingness due to loss to follow-up easily can be incorporated in either method in the presence of ignorable missing data, standard methods are generally invalid when the missingness is non-ignorable or due to death. We propose a Bayesian non-parametric method to simultaneously address both issues. In particular our approach incorporates Bayesian additive regression trees for G-computation estimation of the survivors average causal effect (SACE), i.e. the causal effect on the subpopulation of those surviving irrespective of exposure. The method allows to perform sensitivity analyses for assumptions about missing data mechanisms, i.e. lost to follow-up, death and unmeasured confounders. We illustrate the methodology using longitudinal observational data for studying the effect of widow(er)hood on cognition, where study participants death or drop-out, up until or after the time of the event, are complicating factors.

EO188 Room MAL B20 EXTREME VALUE THEORY AND RISK MODELING Chair: Ivette Gomes

E0394: **Reduced bias kernel value-at-risk estimation**
Presenter: **Frederico Caeiro**, NOVA.ID.FCT - Universidade Nova de Lisboa, Portugal
Co-authors: Ligia Henriques-Rodrigues
The aim is the estimation of the value-at-risk (VaR) at a small level $q \in]0,1[$. The VaR is a key measure in risk market, and represents the size of the loss that occurs with a probability q. For heavy right-tails, the classical VaR estimators are the Weissman-Hill estimators, based on an intermediate number k of top order statistics. Semi-parametric reduced-bias (RB) VaR-estimation procedures based on Kernel estimators of the extreme value index (EVI) are put forward. Under convenient restrictions on the underlying model, these Weissman-Kernel RB VaR-estimators are consistent and asymptotically normal for adequate k, the number of top order statistics to be used. The adequate VaR procedures are then applied to the standardized log-returns of the Bovespa stock market index.

E0891: **Population dynamics: Risk modeling and extremes in the generalized Verhulst model**
Presenter: **Maria Brilhante**, FCiencias.ID, Portugal
Co-authors: Ivette Gomes, Dinis Pestana, Sandra Mendonca
The Verhulst Model $dN(t)/dt = rN(t)[1-N(t)/K]$, for a carrying capacity K, whose logistic solution describes sustainable population growth when the Malthusian reproduction rate r belongs to [1,3], and is proportional to the Rachev and Resnick geo-max-stable logistic law. Observe that this sustainable growth results form the equilibrium between the increase factor $N(t)$ and the retroaction control $1-N(t)/K$. Unstabilities that arise when the Malthusian parameter r falls out of the range [1,3] have been thoroughly described in the chaos literature. The Verhulst model can be looked as an approximation of $dN(t)/dt = r[-ln[1-N(t)/K][lnN(t)]$; other approximations of the form $dN(t)/dt = rN^{p-1}(t)[1-N(t)/K]^{q-1}$ - and more generally of the form $dN(t)/dt = rN^{p-1}(t)[ln[1-N(t)/K]^{P-1}[1-N(t)/K]^{q-1}[-lnN(t)]^{Q-1}$ - for appropriate choices of equilibria of the growth factor and of the retroaction control, have solutions that are proportional to a Rachev and Resnick geo-max-stable law (respectively to a Fisher and Tippett extreme value stable law), either of maxima or of minima. In particular, the solution of $dN(t)/dt = rN(t)[-lnN(t)]$ is the Gompertz function, proportional to the Gumbel extreme value law, and this has been studied as a population growth model of cells in cancer tumors. Our aim is to investigate how far we may use evaluations of growth and of control factors to model risks of extreme events in population dynamics.

E0749: **Insurance portfolio risk minimization**
Presenter: **Alexandra Dias**, University of York, United Kingdom
Co-authors: Isaudin Ismail, Aihua Zhang
The problem of risk minimization of an insurance portfolio from the insurer point of view is considered. We investigate the optimal proportion of different lines of business that minimize a chosen measure of risk. We use extreme value distributions and copulas in order to estimate the measure of risk tail value at risk. Tail value at risk is the measure chosen by Solvency II in order to estimate the solvency capital requirement (SCR) that insurers are required to hold in the European Union. Hence a strategy that minimizes the aggregated risk of the insurance portfolio composed of several lines of business reduces the SCR. We also take into account the use of reinsurance in our empirical study.

E0615: **Extreme M-quantiles as risk measures**
Presenter: **Stephane Girard**, Inria, France
Co-authors: Abdelaati Daouia, Gilles Stupfler
Quantiles are basic tools in extreme-value theory in general, and in actuarial and financial mathematics in particular. The alternative class of expectiles has recently been receiving a lot of attention in actuarial science, econometrics and statistical finance. Both of these notions can be embedded in a more general class of M-quantiles by means of L_p optimization. These generalized L_p quantiles can in some sense, for p between 1 and 2, interpolate between ordinary quantiles and expectiles. We investigate here their estimation from the perspective of extreme values in the class of heavy tailed distributions. We construct estimators of intermediate and extreme L_p quantiles and establish their asymptotic normality in a dependence framework motivated by financial and actuarial applications. We also investigate the potential of extreme L_p quantiles as a tool for

estimating the usual quantiles and expectiles themselves. We show the usefulness of extreme L_p quantiles and elaborate the choice of p through applications to some simulated and financial real data.

E1279: **The effect of global warming in catastrophic risk caracterization**
Presenter: **Isabel Serra**, Centre de Recerca Matematica, Spain
Catastrophic risk is the risk of a large loss by reason of the occurrence of a peril, for instance, the case of the losses for hurricanes, it could result in loss of life, destruction of infrastructure on a large scale. Models for improving adjustments of the destructive power of hurricanes are needed. We analize this models a long the time and show the effect of the global warming. Consequently, we show the effect of global warming in CAT bond pricing for the hurricanes produced in the North Atlantic.

EO429 Room MAL B30 ADVANCES IN STATISTICAL METHODS FOR SURVIVAL ANALYSIS Chair: Marialuisa Restaino

E0502: **Dynamic survival prediction for multivariate joint models using the R package joineRML**
Presenter: **Graeme Hickey**, University of Liverpool, United Kingdom
Co-authors: Pete Philipson, Andrea Jorgensen, Ruwanthi Kolamunnage-Dona
Methods for the joint analysis of time-to-event data and longitudinal data have been developed in recent years, with most emphasis on modelling and estimation. Moreover, research has predominantly concentrated on the joint modelling of a single longitudinal outcome. In clinical practice, the data collected might be more complex, featuring multiple longitudinal outcomes. Harnessing all available measurements in a single model is advantageous and should lead to improved inference and more specific model predictions. In recent years there has been a growing interest in the application of prognostic models to the field of personalised medicine, which can be leveraged by clinicians to adapt care optimally to patients anticipated to deteriorate. Here we focus on the dynamic prediction of a subjects failure time conditional on their observed history of multivariate longitudinal outcome measurements. We will explore the influence of moving from a univariate dynamic prediction to a multivariate framework. We describe an R package – joineRML – recently available on CRAN (https://cran.r-project.org/web/packages/joineRML/index.html), which fits joint models to time-to-event data and multivariate longitudinal data. We demonstrate the latest extension of this package for calculating out-of-data dynamic predictions. The package and methodology are illustrated using a real-world clinical dataset that records several repeatedly measured biomarkers.

E0649: **Penalised competing risks regression**
Presenter: **Federico Ambrogi**, University of Milan, Italy
Co-authors: Thomas Scheike
High dimensional data analysis is an important topic in many research fields. For example, biomedical research generates increasing amount of data to characterise patients bio-profiles (e.g. from genomic, imaging, physiological measurements, laboratory tests etc.). In the last decades many forms of penalized regression have been developed, as a modern form of variable selection, to cope with high dimensional settings. The increasing complexity in the characterisation of patients bio-profiles, is added to the complexity related to the prolonged follow-up of patients with the registration of the occurrence of possible adverse events. Although in the last years the number of contributions for coping with high dimensional data in standard survival analysis have increased, the research regarding competing risks is less developed. The aim is to consider how to do penalized regression when considering the crude cumulative incidence. The direct binomial regression model is reformulated in a penalized framework to possibly fit a sparse regression model. The proposed approach is easily implementable using existing high performance software to do either ridge, or lasso or elastic net penalization. Results from simulation studies are presented together with an application to genomic data when the endpoint is progression free survival.

E1083: **Restricted residual mean lifetime and competing risks**
Presenter: **Giuliana Cortese**, University of Padua, Italy
The typical modeling approach in time-to-event analysis is to consider regression on the hazard function, the main example being the Cox proportional hazards model. Some crucial issues about hazards models have been raised in the recent literature, and currently, there is increasing interest in global summary measures based on the survival function, such as the mean lifetime and the residual mean lifetime. Different regression models for these key measures have been proposed, following either a direct or indirect modeling approach. We present novel regression models based on the residual mean lifetime, and related inference, in presence of right-censoring and left-truncation, and a competing risks structure. We follow a direct approach based on generalized estimating equations, combined with either the IPCW technique or the pseudo-observations technique, to handle incomplete data. In the competing risks setting, the typical key quantity of interest is the cumulative incidence function (CIF). Global summary measures of CIFs would be highly beneficial to provide a clear and direct quantification of between-group differences. We explore different approaches to modeling cause-specific residual mean lifetimes and related measures in presence of competing risks. In particular, we provide a direct competing risks regression model that can handle time-dependent regression coefficients and covariates. The methods and models will be illustrated with applications to clinical data.

E0954: **General discrete non-homogeneous Markov models with multiple absorbing states: Application to breast cancer**
Presenter: **Juan Eloy Ruiz-Castro**, University of Granada, Spain
Co-authors: Mariangela Zenga
A general multi-state non-homogeneous Markov models have been built to analyze the behavior of an illness with several absorbing states. The evolution of a disease occurs in continuous time, but it is observed in discrete time from scheduled revisions or emergency situations. Thus, this analysis has been focused on discrete time. The model is built, covariates depending on time are introduced, the likelihood function for different cases and relevant measures, such as survival function, transition probabilities, mean total times and the conditional probability of state change are determined for different risk groups. Several non-homogeneous Markov models are estimated for analyzing the behavior of breast cancer from a cohort of mastectomized patients. Several discrete probability distributions, such as log-logistic and Weibull, are considered. Cut-points are introduced and they are estimated jointly with the parameters by maximum likelihood. The results are obtained in a matrix algebraic form and they were implemented computationally with MATLAB and R.

E0590: **Restricted mean survival time in time-to-event analysis: From individual level data to aggregate data**
Presenter: **Yinghui Wei**, Plymouth University, United Kingdom
Restricted mean survival time refers to the mean survival time up to a specific time point. As an alternative measure to the widely used hazard ratio, the between-group difference in restricted mean survival time avoids the proportional hazards assumption and allows the effect measure to vary with time. Examples will be given to illustrate the use of restricted mean survival time in individual study. In aggregate data meta-analyses of survival outcomes from multiple studies, an essential step is to extract the estimated effect measures and their variances from publications. However, the relevant information, including restricted mean survival time, may not be reported in publications. We develop a tool to convert the published Kaplan-Meier curves to survival data at individual level, which contain the event indicator as well as the time to event for each individual. The reconstructed data open the possibility to estimate the restricted mean survival time, and hence enable the use of this alternative measure in aggregate meta-analyses of time-to-event outcome.

EO150 Room MAL B33 SEMI- AND NON- PARAMETRIC METHODS FOR FUNCTIONAL STATISTICS Chair: Enea Bongiorno

E0324: **Logratio approach to modeling of densities with application to multivariate functional principal component analysis**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Alessandra Menafoglio, Peter Filzmoser
A concise methodology has been developed since the early 1980s to deal with compositional data - i.e., multivariate data carrying only relative information - through the logratios of their parts. In parallel, the logratio approach to capture the specific features of continuous distributions (densities), known as the Bayes space methodology, is recently developed intensively as well. The aim is is to provide specific details to the continuous case focusing on the implications in case of multivariate functional principal component analysis. The theoretical developments are illustrated with a real-world case study.

E0560: **Local inference for functional-on-scalar mixed models**
Presenter: **Alessia Pini**, Umea University, Sweden
Co-authors: Helle Sorensen, Anders Tolver, Simone Vantini
The problem of performing nonparametric inference on the parameters of a functional-on-scalar mixed effect model is addressed. We perform inference in a local perspective, i.e., defining an adjusted p-value function for each parameter of the model. Such adjusted p-value functions can be thresholded at level alpha to select the regions of the domain presenting statistically significant effects. We show that the p-value functions are provided with an interval-wise control of the family wise error rate. In detail, the probability of wrongly selecting as significant a region of the domain where the null hypothesis is true is always controlled. Since inference is carried out by means of nonparametric permutation tests, the procedure will be exact regardless of the distribution of the functional data, and the sample size. We apply the proposed procedure to test differences between the 3D acceleration curves of trotting horses.

E1022: **Estimation of mean electricity consumption curves for small areas**
Presenter: **Anne de Moliner**, Universite de Bourgogne EDF, France
Co-authors: Herve Cardot, Camelia Goga
The French electricity company EDF is interested in estimating the mean electricity consumption curves of different groups of customers for marketing purposes. These aggregated electricity consumption curves are estimated using samples of thousands of curves selected according to a sampling design and measured at a small time step. Today, there is a growing need for estimations not only at the national level but also for small geographic regions. For some of these regions, there are few units in our panels so the estimators can be very imprecise. This problem of small area estimation in a finite population is very usual in survey sampling and is often addressed by borrowing strength among areas through an explicit modelling of the relation between the auxiliary information and the variable of interest. The goal will be to adapt this existing methodology, developed for totals of real variables, to the framework of functional data. To that purpose two methods are proposed: the first one consists in using a functional principal components analysis in order to transform our functional problem into several uncorrelated real total estimation problems that we can address using standard linear mixed models (known in survey sampling as Unit Level Models). The other method consists in predicting the curve of each non sampled unit using regression trees or random forests adapted to functional data. These methods are tested on real datasets and compared to each other.

E1782: **On the functional Mahalanobis distance**
Presenter: **Beatriz Bueno-Larraz**, Universidad Autonoma de Madrid, Spain
Co-authors: Jose Berrendero, Antonio Cuevas
The theory of Reproducing Kernel Hilbert Spaces (RKHS's) has found many interesting applications in different fields, including statistic. For instance, it helps to partially overcome some difficulties that arise when moving from the multivariate context to the functional one, like the non-invertivility of the covariance operators. One of the problems derived from this non-invertivility is that it does not exist a functional counterpart of the Mahalanobis distance (a relevant notion of multivariate depth). We suggest a suitable functional version of this distance based on the RKHS associated with the underlying stochastic process of the data. This new statistical distance inherits some interesting properties of the original multivariate distance and has shown good performances in different problems (like functional classification, outlier detection, etc).

E1621: **Recovering covariance from functional fragments**
Presenter: **Marie-Helene Descary**, University of Geneva, Switzerland
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 δ-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 observation is discrete, and 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. Our estimator is seen to be consistent in a fully functional sense, and indeed we obtain convergence rates demonstrating that even a parametric rate is attainable provided the grid is sufficiently dense. We illustrate the performance of our method in a real and simulated examples.

EO553 Room MAL B34 RECENT DEVELOPMENTS IN FUNCTIONAL TIME SERIES ANALYSIS Chair: Siegfried Hoermann

E0214: **Detecting and localizing differences in functional time series dynamics: A case study in molecular biophysics**
Presenter: **Shahin Tavakoli**, University of Cambridge, United Kingdom
Co-authors: Victor Panaretos
Motivated by the problem of inferring the molecular dynamics of DNA in solution, and linking them with its base-pair composition, we consider the problem of comparing the dynamics of functional time series (FTS), and of localizing any inferred differences in frequency and along curvelength. The approach we take is one of Fourier analysis, where the complete second-order structure of the FTS is encoded by its spectral density operator, indexed by frequency and curvelength. The comparison is broken down to a hierarchy of stages: at a global level, we compare the spectral density operators of the two FTS, across frequencies and curvelength, based on a Hilbert-Schmidt criterion; then, we localize any differences to specific frequencies; and, finally, we further localize any differences along the length of the random curves, that is, in physical space. A hierarchical multiple testing approach guarantees control of the averaged false discovery rate over the selected frequencies. In this sense, we are able to attribute any differences to distinct dynamic (frequency) and spatial (curvelength) contributions. Our approach is presented and illustrated by means of a case study in molecular biophysics: how can one use molecular dynamics simulations of short strands of DNA to infer their temporal dynamics at the scaling limit, and probe whether these depend on the sequence encoded in these strands?

E0246: **Constrained functional time series: Applications to the Italian gas market**
Presenter: **Antonio Canale**, University of Padua, Italy
Co-authors: Simone Vantini
Motivated by market dynamic modelling in the Italian natural gas balancing platform, a model is proposed to analyze time series of monotone functions subject to an equality and inequality constraint at the two edges of the domain, respectively, such as daily demand and offer curves. In detail, we provide the constrained functions with a suitable pre-Hilbert structure and introduce a useful isometric bijective map associating each possible bounded and monotonic function to an unconstrained one. We introduce a functional-to-functional autoregressive model that is used to forecast future demand/offer functions. We estimate the model via minimization of a penalized mean squared error of prediction with a penalty term based on the Hilbert-Schmidt squared norm of autoregressive lagged operators. The approach is of general interest and is suited for generalizations in any situation in which one has to deal with functions subject to the above constraints which evolve through time.

E0794: **Functional autoregression for sparsely sampled data**
Presenter: **Daniel Kowal**, Rice University, United States
Co-authors: David Matteson, David Ruppert
A hierarchical Gaussian process model is developed for forecasting and inference of functional time series data. Unlike existing methods, our approach is especially suited for sparsely or irregularly sampled curves and for curves sampled with non-negligible measurement error. The latent process is dynamically modeled as a functional autoregression (FAR) with Gaussian process innovations. We propose a fully nonparametric dynamic functional factor model for the dynamic innovation process, with broader applicability and improved computational efficiency over standard Gaussian process models. We prove finite-sample forecasting and interpolation optimality properties of the proposed model, which remain valid with the Gaussian assumption relaxed. An efficient Gibbs sampling algorithm is developed for estimation, inference, and forecasting, with extensions for FAR(p) models with model averaging over the lag p. Extensive simulations demonstrate substantial improvements in forecasting performance and recovery of the autoregressive surface over competing methods, especially under sparse designs. We apply the proposed methods to forecast nominal and real yield curves using daily U.S. data. Real yields are observed more sparsely than nominal yields, yet the proposed methods are highly competitive in both settings.

E0576: **Inference for the autocovariance of a functional time series under conditional heteroscedasticity**
Presenter: **Gregory Rice**, University of Waterloo, Canada
Co-authors: Piotr Kokoszka, Han Lin Shang
Most of the methods for analyzing functional time series rely on the the estimation of lagged autocovariance operators or surfaces. When functional data are constructed from dense records of, for example, asset prices or returns, a weak white noise model allowing for conditional heteroscedasticity is often more realistic. Applying inferential procedures for the autocovariance based on a strong white noise to such data often leads to the erroneous conclusion that the data exhibit significant autocorrelation. We develop methods for performing inference for the lagged autocovariance operators of stationary functional time series that are valid under general conditional heteroscedasticity conditions. These include a Portmanteau test to assess the cumulative significance of empirical autocovariance operators up to a user selected maximum lag, as well as methods for obtaining confidence bands for a functional version of the autocorrelation that are useful in model selection/validation. We analyze the efficacy of these methods through a Monte Carlo simulation study, and apply them to functional time series derived from asset price data of several representative assets. In this application, we found that strong white noise tests often suggest that such series exhibit significant autocorrelation, whereas our tests, which account for functional conditional heteroscedasticity, show that these data are in fact uncorrelated in a function space.

E0591: **Testing for stationarity in functional time series**
Presenter: **Anne van Delft**, Ruhr University Bochum, Germany
Co-authors: Pramita Bagchi, Vaidotas Characiejus, Holger Dette
A new measure for stationarity of a functional time series is proposed, which is based on an explicit representation of the L^2-distance between the spectral density operator of a non-stationary process and its best L^2-approximation by a spectral density operator corresponding to a stationary process. This distance can easily be estimated by sums of Hilbert-Schmidt inner products of periodogram operators (evaluated at different frequencies), and asymptotic normality of an appropriately standardised version of the estimator can be established for the corresponding estimate under the null hypothesis and alternative. As a result, we obtain confidence intervals for the discrepancy of the underlying process from a functional stationary process and a simple asymptotic frequency domain level alpha test (using the quantiles of the normal distribution) for the hypothesis of stationarity of functional time series. Moreover, the new methodology allows also to test precise hypotheses of the form "the functional time series is approximately stationarity", which means that the new measure of stationarity is smaller than a given threshold. Our approach therefore also allows to test for "relevant" deviations from stationarity.

EO449 Room MAL B35 RISK QUANTIFICATION AND EXTREMES Chair: Sebastian Engelke

E0610: **Eigenstructure of sample covariance matrices for high-dimensional heavy-tailed stochastic volatility models**
Presenter: **Thomas Mikosch**, University of Copenhagen, Denmark
Co-authors: Johannes Heiny
The interest is in the asymptotic behavior of the eigenvalues of the sample covariance matrix where the data matrix consists of a p-dimensional time series that constitutes a $p \times n$-dimensional stochastic volatility field. We assume that the marginal tails of the data entries have power-law tails with index smaller than four. We focus on the case when the dimension p increases with the sample size n. Our main goal is to show that the eigenvalues of the sample covariance matrix are essentially determined by its diagonal elements. We consider limit theory of Poisson-type for the point process of the scaled eigenvalues and also discuss the structure of the corresponding eigenvectors.

E1223: **Graphical modelling of extremes**
Presenter: **Adrien Hitz**, University of Oxford, United Kingdom
Co-authors: Sebastian Engelke, Robin Evans
Graphical models provide a useful framework for inferring multivariate distributions in high dimensions by relying on conditional independence. We will explain how these ideas can be applied for modelling multivariate extreme observations. A valid limiting tail distribution must typically satisfy a homogeneity constraint, and we will see how Hammersley-Clifford theorem, a main result in graphical models, can be adapted to this context.

E1208: **A test for Frechet domain of attraction and estimation of the extreme value index**
Presenter: **Andrea Krajina**, University of Goettingen, Germany
An empirical likelihood based test is proposed for a distribution in a Frechet max-domain of attraction and a new estimator of the positive extreme value index. We present the finite sample behaviour and asymptotic properties of both, the test statistic and the estimator.

E1236: **Robust extreme event analysis**
Presenter: **Henry Lam**, University of Michigan, United States

A robust optimization approach is investigated to estimate and quantify the uncertainty of extremal measures of interest from limited data. The approach relies on finding worst-case tail distributions under geometric assumptions and other calibrated auxiliary constraints. We will present some structural results and solution procedures. We will illustrate how the approach balances statistical accuracy with conservatism and connects to conventional extreme value theory.

E1850: **A self-calibrating method for heavy-tailed data modeling: Applications in finance and insurance**
Presenter: **Marie Kratz**, ESSEC Business School, CREAR, France

A new self-calibrating method for heavy-tailed modeling is proposed. One of the main issues in the statistical literature of extremes concerns the tail index estimation, closely linked to the determination of a threshold above which a Generalized Pareto Distribution (GPD) can be fitted. Approaches to this estimation may be classified into two classes, one using threshold methods, in which the threshold to estimate the tail is chosen graphically according to the problem, the other suggesting self-calibrating methods, where the threshold is algorithmically determined. Our approach belongs to this second class proposing a hybrid distribution for heavy tailed data modeling, which links a normal (or lognormal) distribution to a GPD via an exponential distribution that bridges the gap between mean and asymptotic behaviors. A new unsupervised algorithm is then developed for estimating the parameters of this model. The effectiveness of the method is studied in terms of goodness-of-fit on simulated data, and compared with other more standard EVT approaches. One advantage of the method is that it provides a good fit of the entire distribution rather than simply the tail, allowing for a full pricing of the risk when considering applications in risk management. We provide examples of applications in finance, insurance and neuroscience.

EO220 Room CLO 101 ADVANCES IN BAYESIAN METHODOLOGY Chair: David van Dyk

E0472: **Bayesian variable selection under misspecified errors**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
Co-authors: Francisco Javier Rubio

A main challenge in high-dimensional variable selection is enforcing sparsity. Because of theoretical and computational considerations most research are based on linear regression with Normal errors, but in actual applications errors may not be Normal, which can have a particularly marked effect on Bayesian inference. We extend the usual Bayesian variable selection framework to consider more flexible errors that capture asymmetry and heavier-than-normal tails. The error structure is learnt from the data, so that the model automatically reduces to Normal errors when the flexibility is not needed. We show convenient properties (log-likelihood concavity, simple computation) that render the approach practical in high dimensions. Further, although the models are slightly non-regular we show that one can obtain asymptotic sparsity rates under model misspecification. We also shed some light on an important consequence of model misspecification on Bayesian variable selection, namely a potential for a marked drop in power to detect truly active coefficients. This is confirmed in our examples, where we also illustrate computational advantages of inferring the residual distribution from the data.

E1043: **Empirical Bayes model averaging with influential observations**
Presenter: **Christopher Hans**, The Ohio State University, United States

The aim is to investigate the behavior of Bayesian model averaging (BMA) for the normal linear regression model in the presence of influential observations that contribute to model misfit, and to propose remedies to attenuate the potential negative impacts of such observations on inference and prediction. The methodology is motivated by the view that well-behaved residuals and good predictive performance often go hand-in-hand. The focus is on regression models that use variants on Zellner's g prior. By studying the impact of various forms of model misfit on BMA predictions in simple situations we identify prescriptive guidelines for "tuning" Zellner's g prior to obtain optimal predictions. The tuning of the prior distribution is obtained by considering theoretical properties that should be enjoyed by the optimal fits of the various models in the BMA ensemble. The methodology can be thought of as an "Empirical Bayes" approach to modeling, as the data help inform the specification of the prior in an attempt to attenuate the negative impact of model misfit.

E1085: **Using Bayesian visual analytics to conceptualize uncertainty and explore data**
Presenter: **Leanna House**, Virginia Tech, United States

While inundated with big data, uncertainty is also present. There is uncertainty in data collected, uncertainty in methods used to summarize data, and uncertainty in judgements formed from data summaries. Alas, in the presence of big data, data analysts often avoid quantifying uncertainty formally and/or avoid communicating degrees of uncertainty in what might be gleaned from visual or quantitative summaries. There are many potential reasons for avoiding uncertainty, ranging from difficulty (it is hard to model big data well) to misplaced confidence in laws of large numbers to inconsistencies in human behavior. Is it well accepted that, even when measured by probability, the ways by which humans interpret, process, and use uncertainty are personal and vary widely. We take a visual analytic and probabilistic approach to engage humans in learning from big data visually, while considering uncertainty. Specifically, we start with a method we developed called Bayesian Visual Analytics (BaVA) and incorporate novel, visual metaphors of uncertainty, in the context of weighted multi-dimensional scaling visualizations. The door is open to future research into how humans incorporate uncertainty as they visually explore and learn from big data.

E1111: **Interpreting complex models: Efficient, valid posterior inference for meaningful quantities**
Presenter: **Jared Murray**, Carnegie Mellon University, United States
Co-authors: Carlos Carvalho

A conceptually simple framework is proposed for making Bayesian inferences about interpretable models that summarize complex posterior distributions. This provides a vehicle for understanding large, complicated, and often nonparametric models. Our approach is able to map the output of state-of-the-art predictive tools onto scientifically meaningful quantities while maintaining valid posterior inference. It also provides a bridge between Bayesian methods and recently popular frequentist methods for post-selection inference. We illustrate the general approach in two important special cases: Summarizing high-dimensional linear regression models with lower-dimensional alternatives, and interrogating the fit of a nonparametric regression model (Bayesian additive regression trees).

E1172: **Computer model calibration to enable disaggregation of chemical spectra**
Presenter: **David Stenning**, Imperial College London, United Kingdom
Co-authors: Derek Bingham, Kary Myers, Earl Lawrence, Jacob Coleman, Anirban Mondal, Ji Meng Loh, Duane Lee, Robert Wolpert

A novel statistical method is being developed to address a fundamental scientific goal: disaggregation, or estimation of the composition of an unknown aggregate target. By combining computer models of the target of interest with measured data, our approach enables computer-model calibration techniques to directly solve the disaggregation problem. We are developing our method in the context of chemical spectra generated by laser-induced breakdown spectroscopy (LIBS), used by instruments such as ChemCam on the Mars Science Laboratory rover Curiosity. Because a single run of the LIBS computer model may take hours on parallel computing platforms, we build fast emulators for targets that consist of a single chemical compound. These single-compound emulators are combined in a Bayesian hierarchical model for multiple-compound (i.e. aggregate)

targets. We expect our approach to yield the first statistical characterization of matrix effects, i.e. spectral peaks that are amplified or suppressed when chemical compounds are combined in a target versus measured in isolation, and the first capability in uncertainty quantification that addresses the unique challenges of chemical spectra.

EO445 Room CLO 203 CHARACTERIZATIONS OF PROBABILITY DISTRIBUTIONS Chair: Efoevi Angelo Koudou

E0445: **Independencies for Kummer and gamma distributons**
Presenter: **Jacek Wesolowski**, Warsaw University of Technology, Poland
While searching for analogs of the Matsumoto-Yor property, i.e. independence of $X+Y$ and $1/X-1/(X+Y)$ for independent X and Y with generalized inverse Gaussian and gamma distribution, it was previously discovered that for independent X and Y with Kummer and gamma distributions, respectively, random variables $V=X+Y$ and $U=(1+1/(X+Y))/(1+1/X)$ are independent and have Kummer and beta distributions. Similar, but different property was discovered later: if X and Y are independent Kummer and gamma random variables then $V=Y/(1+X)$ and $U=X(1+Y/(1+X))$ are also independent and have Kummer and gamma distributions. We will consider different versions of converse results, some concerned just with independence properties of U and V, some with a weaker assumption of constancy of regressions of U given V. Also a multivariate version of the second property and a related characterization will be presented. It involves a new multivariate version of the Kummer law and tree-based transformations.

E0459: **A Matsumoto-Yor characterization for Kummer and Wishart random matrices**
Presenter: **Bartosz Kolodziejek**, Warsaw University of Technology, Poland
For independent X and Y with Kummer and gamma distributions, respectively, random variables $U=(1+1/(X+Y))/(1+1/X)$ and $V=X+Y$ have been proved to be also independent. This property was later generalized to random matrices X and Y with matrix Kummer and Wishart laws. The aim is to give a converse result in a matrix-variate framework. Such a characterization of matrix Kummer and Wishart laws is proved under the assumption that the densities of X and Y exist and are continuous and strictly positive one some sets. The proof uses the solution to related functional equation with matrix arguments.

E0522: **On transformation of scale distributions on the circle**
Presenter: **Toshihiro Abe**, Nanzan University, Japan
Co-authors: Yoichi Miyata, Takayuki Shiohama
Circular scale distributions on the circle whose densities are asymmetric and unimodal are considered. Basic properties and skewness measures for the distributions are provided. The inverse functions for the skew distributions with mode preserving property are introduced as special cases of this type. Other scale transformations for the distributions are also given. Maximum likelihood estimation for the distributions are also provided as well as the elements of Fisher information matrix. The method of maximum likelihood is applied for a real data set.

E0586: **Modelling earthquakes: Characterizing inter-arrival times and magnitude**
Presenter: **Rosaria Simone**, University of Naples Federico II, Italy
Co-authors: Christophe Ley
Earthquakes modelling is a challenging topic, whose comprehension is needed for the improvement of the understanding of the phenomenon and its dynamics over time and space. After a brief survey on some of the available literature, a preliminary investigation of the probability distributions best-fitting inter-arrival times and strength of the earthquakes is performed. The model search is tailored to identify the desirable properties for the target distributions, entailing effective interpretation of parameters for geo-physical explanation. The ultimate goal is to jointly model the inter-arrival times between subsequent earthquakes and the corresponding magnitude while accounting for location on the sphere within a directional perspective. Data have been downloaded from http://www.ncedc.org/anss/catalog-search.html and refer to the Pacific Ring of Fire for illustrative purposes.

E0928: **A characterization of the Kummer distributions on symmetric matrices**
Presenter: **Pierre Vallois**, Université de Lorraine, France
A characterization of the Kummer distributions of type 2 says that if X,Y_1 and Y_2 are independent random variables such that Y_1 and Y_2 are gamma distributed, with suitable parameters, then $L(X)=L(Y_1/(1+Y_2/(1+X)))$ if, and only if X has the Kummer distribution. The aim is to extend this characterization to the case where X,Y_1 and Y_2 are valued in the cone of symmetric, positive definite real matrices. For the proof, we study the convergence of continued fractions with random matrices entries.

EO634 Room CLO 204 RECENT ADVANCES IN OPTIMAL EXPERIMENTAL DESIGNS Chair: Alexander Donev

E1399: **A review of optimal and efficient designs for choice experiments with partial profiles**
Presenter: **Heiko Grossmann**, Otto-von-Guericke University Magdeburg, Germany
Designs for choice experiments with many attributes require the respondents to process a large amount of information when the alternatives in the choice sets are specified by using all available attributes. The resulting complexity may prompt participants to use simplifying decision rules that violate the additivity assumption on the latent utility scale of the choice model which in turn may invalidate the statistical analysis. One approach to mitigating this problem is to use partial profiles which use only some of the attributes to specify the alternatives in each choice set. Recently, there has been a lot of interest in optimal and efficient designs for this type of choice experiment with contributions from different groups of authors. We present an overview of these developments including some of our own research. We focus on analytic results and corresponding design constructions which at the design stage make the so-called indifference assumption that the utility parameters are equal to zero.

E0358: **New methods for approximating the expected utility in Bayesian design for nonlinear models**
Presenter: **Yiolanda Englezou**, University of Southampton, United Kingdom
Co-authors: David Woods, Tim Waite
The estimation of empirical and physical models is often performed using data collected via experimentation. Hence, the design of the experiment is crucial in determining the quality of the results. For complex models, an optimal design often depends on features, particularly model parameters, which are uncertain prior to experimentation. This dependence leads naturally to a Bayesian approach which can (a) make use of any prior information on these features, and (b) be tailored to the reduction of posterior uncertainty. Optimal Bayesian design for most realistic models is complicated by the need to approximate an analytically intractable expected utility; for example, the expected gain in Shannon information from the prior to posterior distribution. For models which are nonlinear in the uncertain parameters, this expected gain must be approximated numerically. The standard approach employs double-loop Monte Carlo integration using nested sampling from the prior distribution. Although this method is easy to implement, it produces biased approximations and is computationally expensive. We will describe, assess and compare some recent alternatives to simple Monte Carlo sampling from the prior for the approximation of expected utilities. The presented methods include combinations of features from importance sampling and Laplace approximations. Assessments will include both computational cost and the statistical qualities of the resulting approximations.

E0361: **Random designs for misspecified regression models**
Presenter: **Tim Waite**, University of Manchester, United Kingdom
Co-authors: David Woods
A minimax decision theoretic approach to experimental design is considered when it is explicitly acknowledged that a parametric regression model is only an approximation, differing from the true mean response by the addition of a nonparametric discrepancy function. When the class of possible discrepancies is defined by a bound on the L_2-norm, minimax prediction error fails as a selection criterion for finite deterministic designs. This occurs because any deterministic design has infinite maximum risk. However, in other design problems, it has been known since at least the 1980s that it is often minimax optimal to generate design realisations probabilistically according to a randomized strategy. The most familiar example is a simple comparative experiment, in which most commonly there is a fixed and deterministic set of treatments that is allocated randomly to experimental units. We present a novel class of robust random designs for regression in which the treatment set is also generated probabilistically. For these random designs the maximum risk is finite, and sufficiently tractable to enable selection of minimax efficient strategies via numerical optimization algorithms.

E1175: **Comparison of methods for D-optimal design for nonlinear mixed effects models**
Presenter: **Ketil Tvermosegaard**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: John Whittaker, David Woods
Methods to find D-optimal designs for nonlinear mixed effects (NLME) models typically rely on an approximation of the Fisher information matrix (FIM) as a proxy for the variance of the maximum likelihood estimator (MLE). In the classic approach, the model is linearised around the expected value of the random effect and the FIM of this linearised model is used to approximate the true FIM. In a more recent approach, the FIM of the linearised model is viewed as a function of the point around which we linearise and then integrated with respect to the distribution of the random effect. Both these FIM approximations are implemented in generalised R-code and applied to industrially relevant examples drawn from the literature. Clear qualitative differences are established. Resulting designs often differ in terms of location of design points and replication structure. To enable a quantitative comparison of resulting designs, a third, recently proposed, FIM approximation is implemented to act as a gold standard (because it avoids linearisation and theoretically allows for arbitrarily high precision). It is used to compute relative efficiencies of designs. The results indicate that the added integration step most likely does not add to the quality of the resulting designs.

E0223: **Optimum designs for generalised linear models for bivariate response**
Presenter: **Alexander Donev**, University of Manchester, United Kingdom
The development of new products often requires their characteristics to be studied experimentally. When the manufacturing process is prone to gradual changes in the performance of the equipment between calibrations, such changes can be adjusted for by incorporating their effect in the required statistical models for each of the studied characteristics. Different generalised linear models have to be estimated for each of the product characteristics. The construction of a suitable D-optimum experimental design that will be used to collect data for estimating these models requires the use of a composite design criterion. Its choice needs to be done carefully. An example based on an industrial experiment where the response is bivariate will be provided.

EO649 Room MAL 414 STATISTICAL EVALUATION OF MEDICAL DIAGNOSTIC TESTS Chair: Maria del Carmen Pardo

E0252: **The length of the ROC curve as a summary measure under the binormal model**
Presenter: **Maria del Carmen Pardo**, Complutense University of Madrid, Spain
Co-authors: Alba Franco-Pereira, Christos T Nakas
To evaluate the discriminatory ability of a diagnostic marker, it is common to summarize the information of the ROC curve into a single global value or index. The AUC is the most widely used summary measure for the ROC curve in two-class classification problems. It evaluates the overall discriminative power of a diagnostic marker under study. In general, the larger the AUC (when the standard ROC curve is closer to the upper-left corner of the unit square), the higher the distinguishing power of the diagnostic marker. Another summary measure is the maximum of the Youden index defined as the maximum vertical distance between the ROC curve and the chance line (i.e. the main diagonal). We propose the length of the arc of the ROC curve as an alternative index for the evaluation of diagnostic markers in the ROC curve context. We compare the bias and the RMSE of the length index with that of the AUC and the Youden index under the binormal model for the ROC curve under several scenarios. We illustrate with an application on data arising from an applied study.

E0368: **Parametric and non-parametric confidence intervals for the maximum of the Youden index and its associated threshold**
Presenter: **Leonidas Bantis**, The University of Texas MD AndersonCancer Center, United States
Co-authors: Christos T Nakas, Benjamin Reiser
The area under the receiver operating characteristic (ROC) curve as a measure of the overall accuracy of a given biomarker is commonplace. However, after establishing the utility of a biomarker, clinicians in practice need a decision threshold in order to establish whether intervention or simple monitoring is needed. Unnecessary intervention can be avoided and necessary action can be taken based on an optimal decision threshold. The Youden index can be utilized both for the evaluation of the overall accuracy of a biomarker and the derivation of a clinically interpretable optimized decision threshold which can be used for clinical decision making. We present new methods for the construction of parametric and non-parametric confidence intervals for the Youden index as well as its associated decision threshold. Our parametric methods employ the delta approximation and are based on either the normality assumption or a power transformation to normality. Our non-parametric methods are based on kernels and on restricted cubic splines that are forced to be monotone. We compare our approaches to existing ones through simulations and apply them to serum based markers of a prospective observational study involving diagnosis of late-onset sepsis in neonates.

E0632: **Evaluating efficacy of a longitudinal biomarker for clinical endpoint: A joint modelling approach**
Presenter: **Ruwanthi Kolamunnage-Dona**, University of Liverpool, United Kingdom
Joint modelling of longitudinal biomarker and event-time processes has gained its popularity in recent years as they yield more accurate and precise estimates. Considering this modelling framework, a new methodology for evaluating the time-dependent efficacy of a longitudinal biomarker for clinical endpoint is proposed. In particular, the proposed model assesses how well longitudinally repeated measurements of a biomarker over various time periods $(0,t)$ distinguishes between individuals who developed the disease by time t and individuals who remain disease-free beyond time t. The receiver operating characteristic (ROC) curve is used to provide the corresponding efficacy summaries at various t based on the association between longitudinal biomarker trajectory and risk of clinical endpoint prior to each time point. The proposed approach is evaluated through simulation and illustrated on the motivating dataset from a prospective observational study of biomarkers to diagnose the onset of sepsis.

E0642: **Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment**
Presenter: **Cecile Proust-Lima**, INSERM, France
Co-authors: Loic Ferrer
After the diagnosis of a disease, one objective is to predict cumulative probabilities of events such as clinical relapse from the individual information collected up to a prediction time, usually including biomarker repeated measurements. Even before a diagnosis, cumulative probability of disease

can be computed from the individual screening history. Several estimators have been proposed to calculate individual dynamic predictions, mainly from joint modelling and landmarking approaches. These approaches differ by the information used, the model assumptions and the computational complexity. To provide key elements for the development and use of individual dynamic predictions in clinical follow-up, it is essential to properly validate these estimators, quantify their variability and compare them. Motivated by the prediction of two competing causes of prostate cancer progression from the history of prostate-specific antigen, we conducted an in-depth simulation study to validate and compare the dynamic predictions derived from joint models and landmark models. After formally defining the quantity to estimate, we introduce its estimators and propose techniques to assess their uncertainty. We also compare the individual dynamic predictions derived from both approaches in terms of predictive accuracy, efficiency and robustness to model assumptions. We conclude with some recommendations.

E1521: **An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference**
Presenter: **Ariel Alonso Abad**, KU Leuven, Belgium
Co-authors: Wim Van der Elst, Geert Molenberghs

A new metric of surrogacy, the so-called individual causal association (ICA), is introduced using information-theoretic concepts and a causal inference model for a binary surrogate and true endpoint. The ICA has a simple and appealing interpretation in terms of uncertainty reduction and, in some scenarios, it seems to provide a more coherent assessment of the validity of a surrogate than existing measures. The identifiability issues are tackled using a two-step procedure. In the first step, the region of the parametric space of the distribution of the potential outcomes, compatible with the data at hand, is geometrically characterized. Further, in a second step, a Monte Carlo approach is proposed to study the behavior of the ICA on the previous region. The method is illustrated using data from the Collaborative Initial Glaucoma Treatment Study.

EO380 Room MAL 415 RECENT ADVANCES IN TREE-BASED METHODS Chair: Ruoqing Zhu

E0183: **Inference and consistent variable selection for random forests and other tree-based ensembles**
Presenter: **Lucas Mentch**, University of Pittsburgh, United States
Co-authors: Giles Hooker

Despite the success of tree-based learning algorithms (bagging, boosting, random forests), these methods are often seen as prediction-only tools whereby the interpretability and intuition of traditional statistical models is sacrificed for predictive accuracy. We present an overview of recent work that suggests this black-box perspective need not be the case. We begin by developing formal statistical inference procedures for predictions generated by supervised learning ensembles. Ensemble methods based on bootstrapping often improve accuracy and stability, but fail to provide a framework in which distributional results are available. Instead of aggregating full bootstrap samples, we consider a general resampling scheme in which predictions are averaged over trees built on subsamples and demonstrate that the resulting estimator belongs to an extended class of U-statistics. As such, a corresponding central limit theorem is developed allowing for confidence intervals to accompany predictions, as well as formal hypothesis tests for variable significance and additivity. The test statistics proposed can also be extended to produce consistent measures of variable importance that are robust to correlation structures between predictors. Finally, we discuss efficient variance estimation methods for the above procedures and provide demonstrations on ebird citizen science data.

E0355: **Random forests for high dimensional longitudinal data**
Presenter: **Robin Genuer**, Bordeaux University INSERM Vaccine Research Institute, France
Co-authors: Louis Capitaine, Rodolphe Thiebaut

Random Forests are a statistical machine learning method which show good behaviors in high dimensional settings, such as genomic data analysis. However in many problems longitudinal data are available, i.e. measurements are done several times on the same individual-hence observations are not independent-, whereas random forests work on the assumption of i.i.d. samples. Based on semi-parametric mixed models and EM algorithm, we study existing random forests adaptations for high dimensional longitudinal data, as well as a new one. Simulation experiments are done and a real vaccinal trial for HIV dataset, DALIA-1, is analyzed. In this trial, 10 measurements of 32979 gene expressions are available for 18 infected patients. Results show that when the longitudinal aspect of data is taken into account, random forests managed to unravel complex mechanisms between a continuous outcome and a very large number of variables. Furthermore the proposed methodology exhibit faster convergence of EM algorithm and smaller prediction error than existing ones.

E0501: **Bayesian tree ensembles that adapt to smoothness and sparsity**
Presenter: **Antonio Linero**, Florida State University, United States
Co-authors: Yun Yang

Ensembles of decision trees are a useful tool for obtaining for obtaining flexible estimates of regression functions. Examples of these methods include gradient boosted decision trees, random forests, and Bayesian CART. A potential shortcoming of tree ensembles is their lack of smoothness; for example, to approximate a linear function, a single decision tree requires a large number of branches. We show that this problem can be mitigated by instead considering decision trees in which the decisions are treated as probabilistic. We implement this in the context of the Bayesian additive regression trees framework, and show that this approach gives large performance increases on commonly used benchmark problems. We provide theoretical support for our methodology by showing that the posterior concentrates at the minimax rate, up-to a logarithmic factor, for α-Hölder functions, adaptively over α. Our priors over tree structures also allow for adaptation to sparsity; combined with our smoothing of the decision trees, this allows the posterior to concentrate at near-optimal rates adaptively over many classes of functions, including α-Hölder sparse functions and additive functions.

E1322: **Classification trees with mismeasured responses**
Presenter: **Liqun Diao**, University of Waterloo, Canada
Co-authors: Grace Yi

Statistical learning provides a general framework for studying statistical problems including prediction, inference, risk factors identification and risk groups classification using machine learning algorithms, and it has received ever increasing attention in the field of statistics. In the context of statistical learning, potential effects of poor data quality on the performance of machine learning techniques have not received much attention though this problem is important. To close this gap we study the performance of classification trees when the binary responses are not accurately measured, (i.e., mismeasured). Furthermore, we propose a correction algorithm by altering the decision rules for tree building with the mismeasurement of response data taken into consideration; more specifically, we introduce new measures of node impurity and misclassification cost to accommodate response misclassification. We compare the proposed classification trees algorithm to usual classification trees algorithms with response mismeasurement ignored via simulated studies as well as application to real data sets. It is demonstrated that in the presence of response mismeasurement ignoring this feature yields erroneous results and that our proposed statistical learning methods apparently improve upon implementation of unaltered classification trees.

E1393: **Locally linear forests**
Presenter: **Rina Friedberg**, Stanford University, United States
Co-authors: Julie Tibshirani, Susan Athey, Stefan Wager
Random forests are a powerful and ubiquitous predictive method, increasingly considered as an adaptive approach for non-parametric statistical estimation. A limitation of Breiman's forests, however, is that they cannot take advantage of smoothness properties of the underlying signal. This leads to unstable estimates and noticeably suboptimal prediction in the presence of strong linear effects. Local linear regression is another popular method that performs well in this setting, but that traditionally employs non-adaptive kernels and is moreover subject to an acute curse of dimensionality. Drawing on the strengths of these techniques, we introduce locally linear forests, which use an adaptation of Breiman's algorithm to motivate a data-adaptive kernel that can then be plugged into a locally weighted regularized regression. In our experiments, we find that locally linear forests improve upon traditional regression forests in the presence of strong, smooth effects. We contrast our procedure with "model forests", which aggregate within-leaf predictions from many different regression models, and discuss problem-specific splitting rules and inferential methods.

EO224 Room MAL 416 HIGH DIMENSIONAL DATA ANALYSIS WITH APPLICATIONS TO BIOMEDICAL STUDIES Chair: Hongyuan Cao

E1459: **iJRF to study the effect of environmental exposures on miRNA-mRNA interactions in mammary transcript**
Presenter: **Pei Wang**, Icahn School of Medicine at Mount Sinai, United States
Integrative approaches characterizing the interactions among different types of biological molecules have been demonstrated to be useful for revealing informative biological mechanisms. One such example is the interaction between microRNA (miRNA) and messenger RNA (mRNA), whose deregulation may be sensitive to environmental insult leading to altered phenotypes. We introduce a new network approach integrative Joint Random Forest (iJRF), which characterizes the regulatory system between miRNAs and mRNAs using a network model. iJRF is designed to work under the high-dimension low-sample-size regime, and can borrow information across different treatment conditions to achieve more accurate network inference. It also effectively takes into account prior information of miRNA-mRNA regulatory relationships from existing databases. When iJRF is applied to the data from an animal experiment designed to investigate the effect of low-dose environmental chemical exposure on normal mammary gland development in rats, we detected a few important miRNAs that regulated a large number of mRNAs in the control group but not in the exposed groups, suggesting the disruption of miRNA activity due to chemical exposure. Effects of chemical exposure on two affected miRNAs were further validated using breast cancer human cell lines.

E1742: **High dimensional integrative analysis**
Presenter: **Hongyuan Cao**, University of Missouri-Columbia, United States
Large scale multiple testing is a fundamental problem in high dimensional statistical inference. Recent technological advancement makes available various types of auxiliary information such as prior data and external covariates. Our goal is to use such auxiliary information to improve power in a tuning parameter free manner, compared to conventional procedures that only use test statistics or p-values. This is formally achieved through a shape-constrained relationship between auxiliary information and test statistics or p-values. We show that the proposed method leads to a large power increase, while controlling the false discovery rate, both empirically and theoretically. Extensive simulations demonstrate the advantage of the proposed method over several state-of-the-art methods. Dataset from GWAS and multiple-tissue eQTL analysis are used to illustrate this new methodology.

E1674: **A general framework for high-dimensional inference and multiple testing**
Presenter: **Yang Ning**, Cornell University, United States
Co-authors: Han Liu
The problem of how to control the false scientific discovery rate in high-dimensional models is considered. Towards this goal, we focus on the uncertainty assessment for low dimensional components in high-dimensional models. Specifically, we propose a novel decorrelated likelihood based framework to obtain valid p-values for generic penalized M-estimators. Unlike most existing inferential methods which are tailored for individual models, our method provides a general framework for high-dimensional inference and is applicable to a wide variety of applications, including generalized linear models, graphical models, classifications and survival analysis. The proposed method provides optimal tests and confidence intervals. The extensions to general estimating equations are discussed. Finally, we show that the p-values can be combined to control the false discovery rate in multiple hypothesis testing.

E1297: **A multivariate mixed-effects selection model for batch-processed proteomics data with non-ignorable missingness**
Presenter: **Lin Chen**, University of Chicago, United States
In quantitative proteomics, mass tag labeling techniques, such as isobaric tags for relative and absolute quantitation, have been widely adopted in mass spectrometry experiments. These techniques allow peptides/proteins from multiple samples of a batch being quantified in a single experiment, and as such greatly improve the efficiency of protein quantitation. However, the batch-processing of samples also results in severe batch effects and non-ignorable missing data occurring at the batch level. We developed a multivariate MIxed-effects SElection model framework (mvMISE) to jointly analyze multiple correlated genomic features in labeled proteomics data, considering the batch effects and the non-ignorable missingness. We proposed tailored models to account for different correlation structures among specific high-dimensional features. We employed a factor-analytic random effects structure to model the high correlations among multiple peptides, each of which is a shorter fragment digested from the same protein. We introduced a graphical lasso penalty on the error precision matrix for modeling sparse biological dependence among multiple proteins in a functional pathway. We developed estimation algorithms for the models. We applied the proposed methods to the breast cancer proteomic data from the Clinical Proteomic Tumor Analysis Consortium, and identified phosphoproteins/pathways showing differential abundances in triple negative breast tumors versus.

E1477: **Gene regulatory networks for genomewide functional screen data**
Presenter: **Hua Zhong**, NYU School of Medicine, United States
Graphical models are a popular approach to find dependence and conditional independence relationships between gene expressions. Directed acyclic graphs (DAGs) are a special class of directed graphical models, where all the edges are directed edges and contain no directed cycles. The DAGs are well-known models for discovering causal relationships between genes in gene regulatory networks. However, estimating DAGs without assuming known ordering is challenging due to high dimensionality, the acyclic constraints, and the presence of equivalence class from observational data. To overcome these challenges, we propose a two stage adaptive Lasso approach, called NS-DIST,which performs neighborhood selection (NS) in stage 1, and then estimates DAGs by the discrete improving search with Tabu (DIST) algorithm within the selected neighborhood. Simulation studies are presented to demonstrate the effectiveness of the method and its computational efficiency. Two real data examples are used to demonstrate the practical usage of our method for gene regulatory network inference.

EO033 Room MAL 421 NEW ADVANCES IN STATISTICAL DEPTH Chair: Davy Paindaveine

E1148: **Nonparametric imputation by data depth**
Presenter: **Julie Josse**, INRIA, France
Co-authors: Pavlo Mozharovskyi, Francois Husson
A methodology for single imputation of missing values is presented which borrows the idea from data depth. This consists in iterative maximization of the depth of each observation with missing values, and can be employed with any properly defined statistical depth function. On each single iteration, imputation is narrowed down to optimization of quadratic, linear, or quasiconcave function being solved analytically, by linear programming, or the Nelder-Mead method, respectively. Being able to grasp the underlying data topology, the procedure is distribution free, allows to impute close to the data, preserves prediction possibilities different to local imputation methods (k-nearest neighbours, random forest), and has attractive robustness and asymptotic properties under elliptical symmetry. It is shown that its particular case — when using Mahalanobis depth — has direct connection to well known treatments for multivariate normal model, such as iterated regression or regularized PCA. The methodology is extended to the multiple imputation for data stemming from an elliptically symmetric distribution. Simulation and real data studies positively contrast the procedure with existing popular alternatives. The method has been implemented as an R-package.

E0688: **A functional depth example satisfying the axiomatic properties of functional depth**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain
Co-authors: Heather Battey
For making inferences as well as descriptive analyses in functional spaces, it is of vital importance to have a concept of order and centrality equivalent to that given by the real line. This has been studied under the framework of statistical functional depth. We provide an example of depth that satisfies the definition of functional depth. This example is based in a functional concept of symmetry and it is adaptable to different situations through the appropriate selection of a metric space.

E0179: **Halfspace depths for scatter, concentration and shape matrices**
Presenter: **Davy Paindaveine**, Universite libre de Bruxelles, Belgium
Co-authors: Germain Van Bever
Halfspace depth concepts for scatter, concentration and shape matrices are proposed. For scatter matrices, our concept extends a previous one to the non-centered case. Rather than focusing, as in earlier works, on deepest scatter matrices, we thoroughly investigate the properties of the proposed depth and of the corresponding depth regions. We do so under minimal assumptions and, in particular, we do not restrict to elliptical distributions nor to absolutely continuous distributions. Interestingly, fully understanding scatter halfspace depth requires considering different geometries/topologies on the space of scatter matrices. We also discuss the structural properties that a scatter depth should satisfy, and investigate whether or not these are met by the proposed depth. As mentioned above, companion concepts of depth for concentration matrices and shape matrices are also proposed and studied. We illustrate the practical relevance of the proposed concepts by considering a real-data example from finance.

E1710: **Data depth for covariance and spectral density matrices**
Presenter: **Joris Chau**, Universita catholique de Louvain, Belgium
Co-authors: Rainer von Sachs, Hernando Ombao
In multivariate time series analysis, objects of primary interest to study cross-dependences in the time series are the autocovariance matrices in the time domain or spectral density matrices in the frequency domain. Non-degenerate covariance and spectral density matrices are necessarily Hermitian and positive definite. We introduce the concept of a statistical data depth for data observations in the non-Euclidean space of Hermitian positive definite matrices, with in mind the application to collections of observed covariance or spectral density matrices. This allows one to characterize central points or regions of the data, detect outlying observations, but also provides a practical framework for rank-based hypothesis testing in the context of samples of covariance or spectral density matrices. First, the desired properties of a data depth function acting on the space of Hermitian positive definite matrices are introduced. Second, we propose two computationally efficient pointwise and integrated data depth functions that satisfy each of these requirements. Several applications of the new data depth concepts are illustrated by the analysis of multivariate brain signal time series datasets.

E0335: **Expectile depth**
Presenter: **Abdelaati Daouia**, UMR5314 TSE-R CNRS, France
Co-authors: Davy Paindaveine
A new statistical depth function based on the concept of expectiles is introduced. The resulting expectile depth, that is of an L_2-nature, is obtained by replacing quantiles with expectiles in the celebrated halfspace Tukey depth. We derive the theoretical properties of expectile depth and compare them to those of its L_1 antecedent, namely the halfspace Tukey depth, and to those of the only other L_2 depth, namely the zonoid depth. Expectile depth satisfies all depth axioms identified in the literature but provides smoother depth regions than its competitors'. We show that expectile depth computation may be based on a surprising monotonicity property. We introduce companion multivariate expectiles, that are strongly linked to the proposed expectile depth and that behave more naturally than the multivariate expectiles available in the literature. We derive uniform consistency results for the sample versions of the proposed concepts and illustrate our results on some real data examples.

EO130 Room MAL 532 DEPENDENCE MODELS AND COPULAS: THEORY AND APPLICATIONS Chair: Fabrizio Durante

E0287: **Copula bivariate generalized survival models**
Presenter: **Giampiero Marra**, University College London, United Kingdom
Co-authors: Rosalba Radice
Copula bivariate generalized survival models are described, where the parameters of the marginal survivor functions and copula dependence are modelled as functions of additive predictors. The marginals include proportional hazards and proportional odds models. Estimation is carried out using an efficient and stable algorithm that allows for automatic estimation of the smoothing parameters. The models are implemented in the R package SemiParBIVProbit and are illustrated on a medical case study.

E0306: **Generalized additive models for pair-copula constructions**
Presenter: **Thibault Vatter**, Columbia University, United States
Co-authors: Thomas Nagler
Pair-copula constructions are flexible models for the dependence in a random vector that have attracted a lot of interest in recent years. We use generalized additive models to extend pair-copula constructions to allow for effects of covariates on the dependence parameters. We let each pair-copula parameter depend directly on the covariates in a parametric, semi-parametric or non-parametric way. We propose a sequential estimation method that we study by simulation, and apply our method to investigate the time-varying dependence structure between the intraday returns on four major foreign exchange rates.

E1073: **Dissimilarity functions for copula-based hierarchical clustering of continuous variables**
Presenter: **Sebastian Fuchs**, Free University of Bozen-Bolzano, Italy
Co-authors: F Marta L Di Lascio, Fabrizio Durante
A copula–based notion of dissimilarity between continuous random variables is introduced and formalized. Such a concept aims at detecting rank–invariant dependence properties among random variables and, as such, it will be defined as a functional on the collection of all copulas. We show how the provided definition includes previous dissimilarity measures considered in the literature like those derived from measures of association and tail dependence but also those of agglomerative hierarchical type. In the latter case, it turns out that the related clustering procedure does not consider the higher-dimensional dependencies among the involved random variables; for instance, they cannot correctly group variables that are pairwise independent but not globally independent. Finally, we compare novel proposed clustering algorithms (taking into account higher-dimensional dependencies) with classical agglomerative clustering methods.

E1577: **On Sklar's theorem with multivariate marginals**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland
The well-known Sklar's theorem gives a characterization of a family of multivariate distribution functions with given univariate marginals. We will characterize a family of multivariate distribution functions with given multivariate marginals. Furthermore, we will present an algorithm how to construct such distribution functions.

E0795: **Exogenous shock models: Characterization and hierarchical construction**
Presenter: **Henrik Sloot**, Technical University of Munich, Germany
A new characterization theorem for the survival-functions of multivariate failure-times arising in exogenous shock models with non-negative, continuous, and unbounded shocks is presented. These survival-functions are the product of their ordered and individually transformed arguments — where the transformations may depend on the specific order and must fulfill a monotonicity condition. Conversely, every survival-function of that very form can be attained using an exogenous shock model. Furthermore, every extendible exogenous shock model has an alternative stochastic representation as a frailty-model involving some additive subordinator. Finally, examples based on hierarchical additive-frailty models are presented.

EO232 Room MAL 538 CLUSTERING/CLASSIFICATION AND MIXTURES I Chair: Geoffrey McLachlan

E0300: **Advances in robust estimation of skew normal mixtures**
Presenter: **Francesca Greselin**, University of Milano Bicocca, Italy
Co-authors: Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Geoffrey McLachlan
To extend the applicability of classic mixtures to dataset presenting asymmetry, mixtures of skew normal distributions are considered. A robust approach for their estimation allow us to analyze also data affected by outliers or deviations from model assumptions. Some issues related to truncated moment estimation are presented, and methods to overcome their awkward estimation will be discussed, also through applications to real data.

E1126: **Flexible mixtures of factor models based on skew component distributions**
Presenter: **Sharon Lee**, University of Queensland, Australia
Flexible mixtures of skew factor analyzers are gaining increasing attention, being exploited as powerful tools for the modelling, clustering, and dimension reduction of high-dimensional data that exhibit non-normal distributional features. These models have emerged as robust generalizations of the traditional mixtures of factor analyzers, where the assumption of normality for the latent factors is relaxed to cater for skewness in the observed data. We discuss several different formulations of skew factor models and propose to adopt a very flexible form of skew distribution as the density for the component latent factors. This allows the model to accommodate various types of skewness and asymmetry in the data, including multiple arbitrary directions of skewness. As such, it encompasses a number of commonly used models as special cases, such as some versions of the skew normal and skew t-factor analyzers. Parameter estimation can be carried out by maximum likelihood via an EM-type algorithm. The usefulness and potential of the proposed model are demonstrated using both real and simulated datasets.

E0708: **The importance of being clustered: Uncluttering the trends of statistics from 1970 to 2015**
Presenter: **Laura Anderlucci**, University of Bologna, Italy
Co-authors: Angela Montanari, Cinzia Viroli
The recent history of statistics is retraced by analyzing all the papers published in five prestigious statistical journals since 1970, namely: Annals of Statistics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society, series B and Statistical Science. The aim is to construct a kind of "taxonomy" of the statistical papers by organizing and clustering them in main themes. In this sense being identified in a cluster means being important enough to be uncluttered in the vast and interconnected world of the statistical research. Since the main statistical research topics naturally born, evolve or die during time, we will also develop a dynamic clustering strategy, where a group in a time period is allowed to migrate or to merge into different groups in the following one. Results show that statistics is a very dynamic and evolving science, stimulated by the rise of new research questions and types of data.

E1452: **Mixtures of locally-mapped support vector machines**
Presenter: **Hien Nguyen**, La Trobe University, Australia
Co-authors: Geoffrey McLachlan, Geoffrey McLachlan
Support vector machines (SVMs) have been highly successfully in application to classification problems of all sizes. However, the usual construction of SVMs only allow for a singular mapping between input and classification, even if that mapping is nonlinear, such as in the case of kernel SVMs. Via recent approaches to local-mapping of mixture regressions, we construct an approach to SVM classification that allows for different mappings in different parts of the input domain. We demonstrate how our mixtures of locally-mapped SVMs can be estimated via maximum quasi-likelihood estimation and the MM (minorization-maximization) algorithm. Furthermore, we present an online algorithm for its estimation in the face of big data. Some theory regarding the estimator and algorithms are presented, and example applications of the methodology are provided.

E1121: **Distributed and private model-based clustering**
Presenter: **Kaleb Leemaqz**, University of Queensland, Australia
Privacy is becoming increasingly important in collaborative data analysis, especially those involving personal or sensitive information commonly arising from health and commercial settings. The aim of privacy preserving statistical algorithms is to allow inference to be drawn on the joint data without disclosing private data held by each party. We propose a privacy-preserving EM (PPEM) algorithm, a novel scheme for training mixture models for clustering in a privacy-preserving manner. We focus on the case of horizontally distributed data among multiple parties for which cooperative learning is required. More specifically, each party wishes to learn the global parameters of the mixture model while preventing the leakage of party-specific information, including any intermediate results that may potentially be traced to an individual party. Another advantage of PPEM is that it does not involve a trusted third party, unlike most existing schemes that implement a master/slave hierarchy. This helps prevent

information leakage in the case of a corrupted party. For illustration, PPEM is applied to the widely popular Gaussian mixture model (GMM) and its effectiveness is analysed through a security analysis.

EO366 Room MAL 539 NEW ADVANCES FOR COMPLEX DATA Chair: Eric Laber

E1255: **Posterior concentration rates using new empirical priors**
Presenter: **Ryan Martin**, North Carolina State University, United States
In high- and infinite-dimensional problems, Bayesian prior specification can be a challenge. For example, in a high-dimensional regression case, while sparsity considerations drive the choice of prior on the model, there is no genuine prior information available about the coefficients in a given model. Moreover, the choice of prior for the model parameters impacts both the computational and theoretical performance of the posterior. As an alternative, one might be tempted to choose an "informative" empirical prior on the model-specific parameters, depending on data in a suitable way. We will present a new approach for empirical prior specification in high-dimensional problems, based on the idea of centering the prior on a suitable estimator. We will give general conditions that guarantee the corresponding "empirical Bayes" posterior distribution achieves a target rate, even adaptively, and specialize to some examples.

E1210: **Scalable computation with skinny Gibbs sampler for high dimensional Bayesian models**
Presenter: **Naveen Naidu Narisetty**, University of Illinois at Urbana-Champaign, United States
The Bayesian paradigm offers a flexible modeling framework for analyzing data with complex structures, but relative to penalization-based methods, it faces a harsher computational burden due to the posterior computation involved. A particularly challenging problem is to devise scalable Bayesian computational methods for high dimensional data settings that are commonly encountered in many biological applications including gene expression data. We will introduce a new Gibbs sampling algorithm for posterior computation called Skinny Gibbs, which is much more scalable than the standard Gibbs samplers for large datasets. In particular, the complexity of the algorithm is only linear in the number of variables at each iteration. Our Skinny Gibbs algorithm results in the property of strong model selection consistency and is flexible to use in a variety of problems including linear and logistic regressions, and a more challenging problem of censored quantile regression where a non-convex loss function is involved. We will demonstrate the statistical and computational performance of our approach through empirical studies.

E1248: **Inter-modal coupling: A class of measurements for studying local covariance patterns among multiple imaging modalities**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Russell Shinohara, Simon Vandekar
Local cortical coupling was recently introduced as a subject-specific measure for studying localized relationships between cortical thickness and sulcal depth. Although a promising first step towards understanding local covariance patterns that are present between these two specific neuroanatomical measurements, local cortical coupling suffers from a limited scope of imaging modalities that can be analyzed within the framework. We generalize and improve this local coupling measure by proposing an analogue in volumetric space that can be used to produce subject-level patterns of covariation among an arbitrary number of volumetric imaging modalities. Our proposed class of measures, collectively referred to as inter-modal coupling (IMCo), is based on a weighted principal component analysis framework rather than weighted regression. We study IMCo between cerebral blood flow and gray matter density using a sample of youths ages 8-21 from the Philadelphia Neurolodevelopmental Cohort.

E0784: **Functional linear model with dependent regressors in high dimensions**
Presenter: **Cheng Chen**, London School of Economics, United Kingdom
Co-authors: Shaojun Guo, Xinghao Qiao
The functional linear model is one of the most widely used tools of functional data analysis. Existing approaches assume either independent and identically distributed functional regressors or a fixed number of dependent functional regressors. We propose a functional linear models to characterize the linear relationship between a scalar response and high dimensional functional regressors with serial correlation. We develop a penalized least squares approach to perform variable selection for serial correlated functional regressors. We investigate the theoretical properties of our proposed method under mild conditions and illustrate the sample performance through an extensive set of simulation studies and one real world example.

E1468: **Variable selection in sparse additive models**
Presenter: **Cristina Butucea**, University Paris-Est Marne, France
Co-authors: Natalia A Stepanova
The problem of recovery of an unknown multivariate signal of d variables is considered in a sparse additive model, where a smaller number s of variables are actually used. We assume that the additive components are smooth functions of the significant variables. Attempting to reconstruct most, but not all, non-zero components of f, we arrive at the problem of almost full variable selection in high-dimensional regression. For two different choices of a class of smooth functions, we establish conditions under which almost full variable selection is possible,and provide a procedure that achieves this goal. Our procedure is the best possible (in the asymptotically minimax sense) for selecting most non-zero components of f. Moreover, it is adaptive in the parameter s. In addition to that, we obtain an adaptive exact selector for the class of infinitely-smooth functions. Our theoretical results are illustrated with numerical experiments.

EO025 Room MAL 540 SPATIOTEMPORAL DATA ANALYSIS AND ITS APPLICATION Chair: Jian Qing Shi

E0385: **A new clustered temporal point process model with application to social media user data**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States
Co-authors: Yongtao Guan
A new class of clustered temporal point process models is presented to model the posting patterns of users of Sina Weibo, the largest twitter-type online social media in China. The proposed model captures both inhomogeneity in the initial posting time and the clustering pattern in the subsequent posts following the initial post. We develop two EM-type algorithms for estimating the parameters in the proposed model, based on which we cluster the different user patterns. In the application to real data, we discover interesting subgroups of users with distinct behaviors in terms of their initial posts and subsequent posts following the initial posts.

E0580: **Clustering and pairwise multiple comparison of spatial-temporal data based on L1 penalty**
Presenter: **Chi Tim Ng**, Chonnam National University, Korea, South
Co-authors: Woojoo Lee, Youngjo Lee
The pairwise multiple comparison problem can be solved by maximizing a penalized likelihood. This does not only rule out the possibility of drawing contradictory conclusions, but also provides a general method of constructing multiple comparison procedures. Family-wise error rate can be controlled via the tuning parameter of the penalty function. In addition, selection consistency can further be achieved so that all hypotheses are concluded correctly with probability goes to one under certain conditions. Applications to spatial-temporal data in economics and genetics are discussed.

E1703: **Spatio-temporal analysis of seismic data using entropy-based complexity measures in the multifractal domain**
Presenter: **Francisco Javier Esquivel**, University of Granada, Spain
Co-authors: Jose Miguel Angulo, Francisco Javier Alonso

There exist a variety of tools to study the behavior of dynamic phenomena; among others, entropy-based approaches have shown satisfactory results describing structural characteristics of complex systems with scaling properties and long-range interactions. In this context, the two-parameter generalization of the Lpez-Mancini-Calbet complexity measures in terms of Rnyi entropies is widely used. These measures were extended to the multifractal domain based on the increments of the generalized Rnyi dimensions. Following a similar approach, a multifractal relative complexity measure is formulated, in terms of the scaling behavior of Renyi relative entropy in a multifractal system. The spatio-temporal dynamics of seismic series of different nature is studied, in particular based on the relative complexity measure proposed.

E1583: **Spatio-temporal models for georreferenced unemployment data**
Presenter: **Soraia Pereira**, FCiencias.ID, Portugal
Co-authors: Kamil Turkman, Luis Correia, Haavard Rue

Portuguese Labor force surveys, from 4th quarter of 2014 onwards, started geo-referencing the sampling units, namely the dwellings in which the surveys are carried. This opens new possibilities in analyzing and estimating unemployment and its spatial distribution across any region, with particular emphasis on models for point-referenced data. We look at spatio-temporal models that are suitable particularly for estimation and prediction of unemployment in small areas.

E1504: **A Bayesian spatio-temporal model for map reconstruction of remote sensing observations and forest inventory prediction**
Presenter: **Md Kamrul Hasan Khan**, University of Arkansas, United States
Co-authors: Barry T Wilson, Avishek Chakraborty, Giovanni Petris

The USDA Forest Service aims to use satellite imagery for monitoring and predicting change in forest-conditions over the time across large geographic regions within the country. The auxiliary data collected from these imageries, such as brightness, greenness and wetness, are relatively dense in space and time and can be used to efficiently predict how the forest inventory map changes over time. However, it contains a huge proportion of missing values at every location due to practical limitations. Therefore, we develop a spatio-temporal model to reconstruct these missing values from posterior predictive distributions. The model consists of a temporal fixed effect based on periodic patterns and a spatio-temporal random effect based on conditional autoregressive (CAR) prior. To allow for presence of change points in the landscape (that should prevent spatial smoothing), we modify the neighborhood structure using binary boundary parameters with a Markov prior over time. Once we obtain full spatio-temporal map, we use it to model the presence/absence of forest and the amount of basal area across the region. These models are formulated using functional regression and elastic net regularization is performed to identify important auxiliary variables.

EO571 Room MAL 541 RECENT RESULTS IN CHANGE POINT ANALYSIS OF MULTIVARIATE DATA Chair: Zuzana Praskova

E0213: **Bootstrap in Hilbert spaces and the detection of changes in distribution**
Presenter: **Martin Wendler**, University of Greifswald, Germany

The Cramer-von Mises-statistic is widely used for testing hypothesis on distribution of random variables. We propose to use it in a CUSUM type test statistic to detect changes in the marginal distribution function of time series or random fields. This can be rewritten as a change point problem with functional observations. The asymptotic distribution of the test statistics is obtained with the continuous mapping theorem from new functional central limit theorems for the partial sum process in a Hilbert space. Because the limit distribution is difficult to evaluate and depends on a high-dimensional, difficult to estimate variance parameter, we propose to use bootstrap methods. In the case of time series, we study the non-overlapping block bootstrap, in the case of random fields the dependent wild bootstrap and show the validity of these methods. In a simulation study, we will show that our test has similar performance as the classical CUSUM test in the case of a mean shift, but better power if the distribution changes in a different aspect.

E0285: **Properties of change point estimates for short time series with missing data**
Presenter: **Daniela Jaruskova**, Czech Technical University, Faculty of Civil Engineering, Czech Republic

A time series of independent normally distributed random variables or vectors is observed and its mean shifts from one unknown value to a different unknown value at an unknown time point. For a confidence intervals construction an asymptotic distribution is usually applied. There exist two formulations of the problem. In the first, the change point is an integer and its estimator maximizes a certain two-sided random walk. In the second, the change point is modelled as a fraction of time and the size of the change tends to zero at a certain rate. Sometimes in applications the studied series are relatively short. The first question may be which of these two approaches is more accurate. Moreover, in multiple time series analysis we often encounter a situation where in several components the data are missing while they are present in the remaining components. One may ask whether it is better to cut off a time interval where the data are missing or to replace them by a constant, e.g. by an average. The answers obtained from a simulation study are rather unexpected.

E0497: **Change point detection in VAR models based on characteristic functions**
Presenter: **Zdenek Hlavka**, Charles University, Czech Republic
Co-authors: Marie Huskova, Simos Meintanis

A new method for detecting structural breaks in multivariate time series (VAR models) is proposed using L2-type criteria based on the empirical characteristic function (CF). The advantage of using CF-based procedure is that vector observations are linearly projected onto the real line and the resulting statistic may be written in a convenient closed-form expression. Asymptotic as well as Monte-Carlo results are presented. The new method is applied to time-series data from the financial sector.

E0226: **Change point analysis in panel data without boundary issue**
Presenter: **Michal Pesta**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Barbora Pestova

Panel data of our interest consist of a moderate number of panels, while the panels contain a small number of observations. An estimator of common breaks in panel means without a boundary issue for this kind of scenario is proposed. In particular, the novel estimator is able to detect a common break point even when the change happens immediately after the first time point or just before the last observation period. Another advantage of the elaborated change point estimator is that it results in the last observation in situations with no structural breaks. The consistency of the change point estimator in panel data is established. The results are illustrated through a simulation study. As a by-product of the developed estimation technique, a theoretical utilization for correlation structure estimation, hypothesis testing, and bootstrapping in panel data is demonstrated. A practical application to non-life insurance is presented as well.

E0798: **Sequential break detection in panel data**
Presenter: **Zuzana Praskova**, Charles University, Czech Republic
A panel data model with lagged dependent variables and unobserved individual effects is considered and a sequential procedure to detect change in coefficients of lagged variables is proposed. The test statistic to detect change is based on quasi-likelihood scores and quasi-maximum likelihood estimators computed from a training data set. Asymptotic properties of the test statistic are studied in case that both the number of panels and the number of observations in the training set are sufficiently large.

EC695 Room MAL 402 CONTRIBUTIONS IN ROBUST STATISTICS Chair: Peter Rousseeuw

E1709: **Principal component analysis of compositional tables using classical and robust methods**
Presenter: **Julie Rendlova**, Palacky University, Czech Republic
Co-authors: Kamila Facevicova, Karel Hron, Peter Filzmoser
Many practical examples contain relative information about the distribution according to two factors which leads to a $(I \times J)$-dimensional extension of compositional data carrying information about a relationship between these factors. Such a structure, called a compositional table x, can be decomposed orthogonally into its independent and interactive parts within the logratio methodology in order to provide a better insight to the original data structure. One of the primary tasks in multivariate statistics is to reduce the dimensionality of the data at hand, done using principal component analysis. To weaken the influence of outliers in PCA, the covariance matrix for robust PCA might be estimated by the Minimum Covariance Determinant estimator. Since popular clr coefficients lead to singularity and are generally not appropriate for robust methods, loadings and scores for PCA need to be computed from pivot coordinates of the interaction and independence tables and then transformed back to clr coefficients for better interpretation of the resulting compositional biplot. Accordingly, the aim is to propose a robust approach to principal component analysis of compositional tables and to illustrate the theoretical background on a real data set from OECD Statistics.

E1707: **L-moments in linear regression model**
Presenter: **Martin Schindler**, Technical University of Liberec, Czech Republic
Co-authors: Jan Picek
The L-moments are analogues of the conventional moments and have similar interpretations. They are calculated using linear combinations of the expectation of ordered data. Earlier it was shown that L-moments are special case of L-estimators. We propose a generalization of L-moments in the linear regression model based on regression quantiles as special L-estimator. The properties of extended L-moments are illustrated on simulated data.

E1754: **Robust regression on compositional variables including cell-wise outliers**
Presenter: **Nikola Stefelova**, Palacky University Olomouc, Faculty of Science, Czech Republic
Co-authors: Andreas Alfons, Javier Palarea-Albaladejo, Peter Filzmoser, Karel Hron
Multivariate data are commonly arranged as a rectangular matrix with observations or cases in the rows and variables in the columns. Ordinary robust estimators are designed to deal with case-wise outliers, assuming that most observations are free of contamination. However, this approach may lead to a significant loss of information in situations where outliers, not necessarily many, occur at the individual cell level but affect a large fraction of observations. Moreover, additional problems are confronted when data of compositional nature are involved. In this case, all the relevant information for statistical analysis is contained in the ratios between parts of the composition, columns of the data matrix, and then cell-wise contamination in these easily propagates throughout and distorts the results. The aim is to present a method for robust compositional regression that is able to deal with both case-wise and cell-wise outliers. In brief, outlying cells are firstly filtered out and replaced by sensible values. Then, robust compositional MM-regression is carried out on the replaced dataset. Imputation uncertainty is reflected on regression coefficient estimates via a multiple imputation (MI) scheme. The performance is assessed using simulated as well as real-world biological data.

E1812: **Robust estimation of multilevel models with sufficient sample sizes**
Presenter: **Ronrick Da-ano**, University of the Philippines-Diliman, Philippines
The aim is to estimate multilevel models with sufficient sample sizes with structural change using a hybrid of forward search algorithm preceding bootstrap method. The influence of the different sample sizes at the group level on the accuracy of the estimates and its standard error was examined. The simulation study exhibits a comparable predictive ability with the classical restricted maximum likelihood estimation (REML). However, the hybrid method yields estimates of the parameters with lower bias in respect to REML. The hybrid of forward search algorithm preceding bootstrap method can even provide more robust estimates of fixed and random coefficients under various levels of interclass correlation or in the presence of structural change.

E1632: **Robust estimation of treatment effects in a latent-variable framework**
Presenter: **Mikhail Zhelonkin**, Erasmus University Rotterdam, Netherlands
The policy evaluation is one of the central problems in modern economics. Unfortunately, it is often impossible to perform a randomized experiments in order to evaluate the treatment effects. Hence, the data from observational studies has to be used. In this case the sample is typically non-random and one has either to correct for selectivity or to impose (conditional) independence assumption. Since this assumption is often unrealistic, the structural latent variable model is used. The parametric estimators (although, they are straightforward to compute and to interpret) have been criticized for sensitivity to the departures from the distributional assumptions. The alternative semi- and non-parametric estimators have complex identification and are limited to estimation of certain parameter(s) of interest but do not allow for the general evaluation and interpretation of the model. We employ a previous latent-variable framework. We study the robustness properties of the estimators of four principal parameters (average treatment effect, average treatment effect on the treated, local average treatment effect and marginal treatment effect), and propose the robust alternatives.

EP002 Room Macmillan Hall and Crush Hall POSTER SESSION Chair: Marios Fyrillas

E0266: **Linking climate and incidence of dengue in the Philippines with a dynamic spatio-temporal model**
Presenter: **Stephen Jun Villejo**, University of the Philippines, Philippines
Co-authors: Rutcher Lacaza
The effect of climate indicators on epidemics can be confounded with site-specific factors. Modeling of complex systems like the spread of epidemics requires a more adaptive and flexible structure like a dynamic spatio-temporal model proposed. The proposed model is superior over ordinary regression with auto-correlated errors. The model provided additional insights on the distribution of the residuals for space-time points with significant random effects. The tail distribution of the significant group has more variability, which also contains the extreme values. This is especially prominent for rainfall, dew point temperature, and vapor pressure which are important indicators for dengue-carrying vector abundance and population dynamics of mosquitoes.

E0404: **Continuous glucose monitoring using distributional regression models**
Presenter: **Jenifer Espasandin-Dominguez**, University of Santiago de Compostela, Spain
Co-authors: Carmen Cadarso Suarez, Thomas Kneib, Francisco Gude

The technological progress has led to the development of new measurement procedures in the form of functional data. We propose to incorporate this functional information within the framework of the distributional regression models. This type of models are a generic framework for performing regression analyses where every parameter of a potentially complex response distribution is related to an additive predictor. In the Bayesian inference framework, structured additive distributional regression models extend the use of generalized additive models to situations in which the response distributions are nonstandard, and where not only the mean but multiple parameters are related to additive predictors. Further, they allow additional flexibility by specifying structured additive predictors for each parameter of interest, and thus adjust for several types of covariate effects. The methodologies developed will be applied to real biomedical data, in a study of glycated haemoglobin, a test useful in the control of individuals with diabetes. The predictor will include the results of continuous monitoring, which collects glucose measurements every 5 minutes over a week. The glucose levels will be included as a functional covariate in a scalar on functional regression model. The inclusion of glucose profiles as a predictor of this type of models will mark a novel advance in the study of diabetes.

E0359: **Bivariate copula regression models in cardiovascular disease**
Presenter: **Pedro Oliveira**, EPIUnit-ICBAS-Universidade do Porto, Portugal
Co-authors: Jenifer Espasandin-Dominguez, Isabel Vila, Jorge Cotter, Pedro Cunha, Carmen Cadarso Suarez

Portuguese population presents a high incidence of stroke that seems to be linked to hypertension, salt consumption, excess weight and other cardiovascular risks. Early vascular aging (EVA) is defined as an accelerated aging process of the blood vessels by which arteries present the characteristics of older chronological ages. It is envisaged that the effects of accumulated cardiovascular risks in vascular aging are reflected in Pulse Wave Velocity and Central Blood Pressure. The data consists of a random sample of dwellers, aged 18-96 years, in Northern Portugal. With the aim of studying the relationship between pulse wave velocity and central blood pressure, depending on gender, age, waist circunference, salt consumption and other clinical covariables, bivariate copula additive models for location, scale and shape for continuous responses will be used. This type of model extends the use of generalized additive models for location, scale and shape to situations in which two responses are modeled simultaneously conditional on some covariables using copulae. The approach permits the two responses and copula parameter to be modelled using additive predictors that allow for several types of covariate effects. The models can be easily used via the SemiParBIVProbit() function in the R package SemiParBIVProbit.

E0448: **Goodness-of-fit tests for regression models**
Presenter: **Sandie Ferrigno**, University Nancy Lorraine/INRIA Nancy, France
Co-authors: Marie-Jose Martinez, Romain Azais

The objective is to construct a tool to test the validity of a regression model. For this, we have studied some omnibus tests of goodness-of-fit. The focus is on the regression function in the regression model. These tests can be "directional" in that they are designed to detect departures from mainly one given assumption of the model (regression function, variance function or error) or global with the conditional distribution function. We have compared, through simulations, different nonparametric methods to test the validity of the model. Two methods are directional and one is global. The establishment of such statistical tests require nonparametric estimators and the use of wild bootstrap methods for the simulations.

E0514: **A semi-widely linear prediction algorithm for stationary Gaussian quaternion signals**
Presenter: **Rosa Maria Fernandez-Alcala**, University of Jaen, Spain
Co-authors: Jesus Navarro-Moreno , Juan Carlos Ruiz-Molina, Jose Domingo Jimenez-Lopez

The prediction problem in the quaternion domain is addressed under stationarity, Gaussianity and C-properness hypotheses. A very general formulation of the problem is considered allowing the estimation of any linear or nonlinear functional of the signal. In the proposed methodology, based on a semi-widely linear (SWL) processing, we first consider a SWL version of the Durbin-Levinson algorithm for the one-stage prediction problem. Then, a general recursive prediction algorithm is devised by incorporating the information supplied by the square vector of the observation quaternion process. The proposed solution presents the following advantages in relation to the conventional quaternion widely linear predictor: better performance, matrices with lower dimensions and efficiently implementable. The practical application of this SWL prediction algorithm is illustrated by means of a numerical example where a nonlinear functional of the signal is estimated.

E1392: **A simulation-study on pseudo likelihood estimation for item response models of binary data under uncertainty**
Presenter: **Mia Johanna Katharina Mueller-Platz**, RWTH Aachen University, Germany
Co-authors: Maria Kateri, Irini Moustaki

Latent variables are widely used to measure constructs that are not directly observed such as attitudes, intelligence, abilities, etc. The latent variables are measured using multiple observed indicators often of categorical nature (ordinal, binary, nominal). A full information maximum likelihood approach for model fitting is based on the marginal distribution of the response vector, integrating out the latent variables. In this way, the distributional assumptions made for the latent variables introduce a further source of uncertainty and possible misspecification. Moreover, multidimensional latent variable models require the approximation of multidimensional integrals and often become intractable. An attractive alternative is the pseudo likelihood estimation procedure, which tackles the model fitting with a flexible and feasible approximation of the marginal distribution. This procedure is adopted in the context of IRT models for binary responses, targeting at developing a fast model fitting technique, appropriate for settings with many items and many latent variables. A simulation-study is used to study the robustness of the proposed procedure against model misspecification and non-normal latent variable distributions.

E1501: **Robust recursive estimation of GARCH models**
Presenter: **Radek Hendrych**, Charles University, Czech Republic
Co-authors: Tomas Cipra

The generalized autoregressive conditional heteroscedasticity (GARCH) process is a particular modelling scheme, which is capable of forecasting the current level of volatility of financial time series. This classic benchmark model is designed to track changes in volatility of financial returns by using past squared measurements. Recently, recursive estimation methods suitable for this class of stochastic processes have been introduced in the literature. They undoubtedly represent attractive alternatives to the standard non-recursive estimation procedures with many practical applications. It might be truly advantageous to adopt numerically effective estimation techniques that can estimate and control such models in real time. However, abnormal observations (outliers) may occur in data. They may be caused by many reasons, e.g. by additive errors, measurement failures or management actions. Exceptional data points will influence the model estimation considerably if no specific action is taken. The aim is to propose and thoroughly examine a robust recursive estimation algorithm suitable for GARCH models. Monte Carlo experiments are performed in order to investigate the introduced approach. Moreover, real data examples are also discussed.

E1505: **Optimum periodic geometries that maximize the heat transfer rate from a solid slab**
Presenter: **Marios Fyrillas**, Frederick University, Cyprus

The inverse problem of determining the optimal geometries/shapes that would maximize the heat conduction rate from a slab is considered. The algorithm will be developed through a combination of conformal mapping techniques (Generalized Schwarz-Christoffel transformation) and Boundary Element methods. The Shape Optimization problems are posed as nonlinear programming problems (constrained nonlinear optimization) where the Objective Function is the Heat Transfer Rate, or the Shape Factor, and the variables of the optimization are the parameters of the Generalized Schwarz-Christoffel transformation. The Shape Optimization problems are addressed through numerical optimization which can handle complicated geometrical constraints. The above methodology has been effectively applied to obtain the optimum geometries of embedded pipes and extended surfaces/fins.

E1509: **Multiple-use confidence regions in multivariate calibration problem**
Presenter: **Martina Chvostekova**, Institute of Measurement Science, Slovak Academy of Sciences, Slovakia

The problem of statistical multivariate calibration is considered in the setup where a normally distributed response variable is related to an explanatory variable through a multivariate linear regression model. The statistical multivariate calibration problem consists of constructing region estimates for future unobserved values of m-dimensional explanatory variable corresponding to possibly infinitely many future observations of q-dimensional response variable by repeatedly using a calibration data from a calibration experiment. The region estimates are referred as multiple-use confidence regions. It is required that at least γ proportion of the constructed multiple-use confidence regions after observing a sequence of future responses will contain the corresponding true value of an explanatory variable. Since the multiple-use confidence regions are derived using the same calibration data, the probability of a γ coverage is $1-\alpha$. We present a procedure for determining the multiple-use confidence regions employing the tolerance regions. The provided numerical investigation shows that the regions so obtained satisfy the coverage requirements quite well. The computational aspects and the practical implementation of our multiple-use confidence regions and those previously published are illustrated using an example.

E1555: **A new approach for ARMA order estimation based on clipping**
Presenter: **Samuel Flimmel**, University of Economics in Prague, Czech Republic
Co-authors: Jiri Prochazka

Problems related to big data are faced in many fields nowadays. Enormous amounts of information are being stored every second, but processing all these data using standard methods becomes more and more problematic. With an increasing number of observations, the probability of outlier presence also rises, and, therefore, working with sufficiently robust methods gains on importance as well. Since standard methods are not always able to deal with outliers correctly, standard estimates are often biased. ARMA processes are well known and widely used in theoretical and also practical world of statistical modeling. Often, it is necessary to estimate the order of a given ARMA process. Usually, it is the second step in Box-Jenkins method, performed immediately after solving the stationarity and seasonality. In this poster, we present a new robust method for ARMA order estimation. The method is based on clipping the original time series and working with a binary time series instead. This detour provides the required robustness and helps to face outliers. We describe this new method and compare it with existing methods using a simulation study performed in the R statistical software.

E1553: **Symbolic data analysis using classification tree**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Masataka Kuroda, Etsuo Miyaoka

The focus is on the construction method of classification trees based on multi-valued covariates, which are given as elements of the power set of the covariate space. When the covariates are given as multi-values, it is not possible to apply the classical splitting rules directly for constructing a classification tree. Moreover, it is important how to use the distribution information of individual descriptions in the covariate space of each sample for the construction of the model. To address these problem, we propose a new splitting rule, and examine its performance through simulation studies. In addition, we present an application of this model in reference to the classification problem of molecular data which are represented by atom-pair properties.

E1569: **Convex variable selection for high-dimensional linear mixed models**
Presenter: **Jozef Jakubik**, Institute of Measurement Science, Slovak Academy of Sciences, Slovakia

Analysis of high-dimensional data is currently a current field of research, thanks to many applications e.g. in genetics (DNA data in genome-wide association studies), spectrometry or web analysis. The type of problems that tend to arise in genetics can often be modelled using high-dimensional linear mixed models because linear mixed models allow us to specify the covariance structure of the models. This enables us to capture relationships in data such as the population structure, family relatedness, etc. The high-dimensional setting presents specific theoretical as well as computational challenges. For high-dimensional linear mixed models there exist a few approaches based on ℓ_1 penalization or SCAD. These methods lead in general to non-convex problems. We present a convex approach to variable selection in high-dimensional linear mixed models for data with dimension over 10^5, where current non-convex approaches often fail. Our method provably ensures consistent variable selection with a growing number of observations. The method achieves good results in experiments on synthetic as well as real data.

E1576: **Application of transition models for higher-order crossover designs with binary responses**
Presenter: **Takuma Kurosawa**, Tokyo University of Science, Japan
Co-authors: Asanao Shimokawa, Etsuo Miyaoka

Higher-order crossover designs are considered, which include more than two sequences or more than two periods or both. The simplest crossover design, AB/BA design without baselines, has some disadvantages that test for an effect of carryover or direct by period interaction lacks power and it alias the effect of carryover, direct by period interaction and group difference each other. We able to use higher-order crossover designs to avoid these undesirable characteristics. These designs, however, tend to be complicated and be larger designs, which needs a lot of sequences and periods that is why needs much more subjects and long terms like orthogonal Latin square designs. To address this problem, we apply transition model to higher-order crossover designs with binary responses and test efficiency of the model through some simulation studies. Moreover, we show the results of applying the model to actual data.

E1657: **Two stage rule to control two-dimensional false discovery rate for composite null hypothesis**
Presenter: **Youngrae Kim**, Seoul National University, Korea, Korea, South
Co-authors: Johan Lim, Jaesik Jeong, Jong Soo Lee

The mass spectral data are featured as being high dimensional with small number of signals (peaks) and many noisy observations. This unique features of mass spectral data motivates the problem testing of many composite null hypotheses simultaneously. The composite null we consider is the union of two simple hypotheses (e.g., the alternative hypothesis is means of two populations are not same and the variance is not too small so that the observations are from signal). We develop two stage procedure to control the false discovery rate of the simultaneous multiple hypothesis testing. The new procedure provides an explicit boundary in two dimensional space of bivariate test statistics, compared to the existing local false discovery rate (fdr) based procedure. We numerically compare the performance of our procedure to the existing, two-dimensional local false

discovery rate (2d-fdr) under various setting, and apply the procedure to differentiating origins of herbal medicine using gas chromatography-mass spectrometry (GC-MS). Finally, we additionally discuss two stage additive procedure for testing many intersection null hypotheses simultaneously.

E1696: **Simulation of major heat waves in central Europe in EURO-CORDEX RCMs**
Presenter: **Jan Kysely**, Institute of Atmospheric Physics AS CR, Czech Republic
The aim is to evaluate the capability of EURO-CORDEX regional climate models (RCMs) to simulate major heat waves in Central Europe and their associated meteorological factors. Three reference major heat waves (1994, 2006, and 2015) were identified in the observed gridded data set (E-OBS), based on their temperature characteristics, length and spatial extent. Atmospheric circulation, precipitation, net short-wave radiation, and evaporative fraction anomalies during these events were assessed using the ERA-Interim reanalysis. Analogous major heat waves and their links to the aforementioned factors were analysed in an ensemble of EURO-CORDEX RCMs driven by various global climate models in the 1970-2016 period. All three reference major heat waves were associated with favourable circulation conditions, precipitation deficit, reduced evaporative fraction and increased net short-wave radiation. This joint contribution of large-scale circulation and land-atmosphere interactions is simulated with difficulties in majority of the RCMs, which affects the magnitude of modelled major heat waves. The findings point to different driving mechanisms of major heat waves in some RCMs compared to observations, which should be taken into account when analysing and interpreting future projections of these events.

E1750: **Mendel mixtures of genuine and fake p-values in meta analysis**
Presenter: **Dinis Pestana**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal
Co-authors: Maria Brilhante, Sandra Mendonca
Meta-analysing p-values is simple under the naive and optimistic assumption that the null hypothesis H_0 is true in each of the independent tests performed, i.e. that the reported p-values come from independent standard uniforms. Assuming that H_0 holds in all cases is farfetched; on the other hand, one of the sources of bad science is the repetition of experiments in order to report a more convenient p-value. Fisher suggested that Mendel used such 'fake' p-values, so that assuming that H_0 holds some of the p-values are genuine (uniforms), but others are fake, namely a Beta(1,k) or a Beta(k,1) extremes of uniforms. It seems therefore sensible to assume that the appropriate model for the sequence of p-values, if H_0 is true, is a mixture of uniforms and extremes of uniforms, that in the sequel is named Mendel model. We present results on the estimation of the mixing parameter, and discuss auto-regressive versus independent setups.

E1687: **Exploring fire incidence in Portugal using generalized additive models for location, scale and shape**
Presenter: **Antonia Turkman**, FCiencias.ID, Portugal
Co-authors: Ana Sa
Wildfires are responsible for large environmental, ecological and socio-economic impacts. Portugal has a +40-year fire atlas of burnt perimeters, allowing the assessment of the fire most prone areas. Burnt area (1975-2009), topography, climate, vegetation and population data were compiled into a 2km grid cell size overlaying mainland Portugal. Fire incidence (FI), defined as the ratio between the total burned area during the 35-year period and the area of each cell, was modelled as a function of those predictors, using generalized additive models for location, scale and shape. By exploring the drivers of fire incidence dispersion and asymmetry parameters it is expected to further understand the forces that promote regional fire incidence and thus to contribute with relevant information for fire modelers and fire management decisions.

E1607: **CharFunTool: The characteristic functions toolbox**
Presenter: **Viktor Witkovsky**, Slovak Academy of Sciences, Slovakia
CharFunTool is a MATLAB repository of characteristic functions and tools for their combinations and numerical inversion.The Characteristic Functions Toolbox (CharFunTool) consists of a set of algorithms for evaluating selected characteristic functions and algorithms for numerical inversion of the combined and/or compound characteristic functions, used to evaluate the cumulative distribution function (CDF), the probability density function (PDF), and/or the quantile function (QF).The toolbox comprises different inversion algorithms, including those based on simple trapezoidal quadrature rule for computing the integrals defined by the Gil-Pelaez formulae, and/or based on using the FFT algorithm for computing the Fourier transform integrals.We shall present basic functionality of the toolbox and illustrate its usage for selected statistical applications. For current status of the MATLAB toolbox see the CharFunTool development available at https://github.com/witkovsky/CharFunTool.

E1836: **Predicting survival risks in the presence of multiple longitudinal markers: A simulation study**
Presenter: **Ayca Pamukcu**, Acibadem University, Turkey
Co-authors: Ozgur Asar
In prospective medical studies, typically baseline covariates, multiple longitudinal biomarkers and time to a survival event are collected. Available literature has emphasized the importance of using repeated biomarker measurements for predicting survival outcomes. Nonetheless, majority has considered the use of a single longitudinal biomarker. We aim to investigate the predictive ability of more than one biomarker. We conduct extensive simulation studies to inspect this contribution under several scenarios, e.g. under different correlation structure between multivariate markers, and different associations between markers and risk of survival event.

E1837: **Fuzzy mortality models based on Von Neumann algebras**
Presenter: **Agnieszka Rossa**, University of Lodz, Poland
Co-authors: Andrzej Szymanski
The generalization of the well-known Lee-Carter mortality model defined in terms of fuzzy numbers was first introduced by proposing a fuzzy representation of the log-central mortality rates and models parameters, by converting them into fuzzy numbers with triangular symmetric membership functions. In our approach we used exponential membership functions decomposed into two parts: strictly decreasing and strictly increasing functions, and transformed into the complex functions of the general form: $f(u) + ig(u)$. The pair of ordered complex numbers is called a quaternion. It can be easily seen that quaternion algebra with a proper defined definition of algebra multiplication is a $C*$-Banach algebra, which is non-commutative. A von Neumann algebra was previously applied as a $C*$-algebra of projections in a separable complex Hilbert space. We applied this approach to define a new mortality model based on von Neumann algebra.

E1845: **Between-team competition, within-team cooperation and team composition: An experimental study**
Presenter: **Francisca Jimenez-Jimenez**, Universidad de Jaen, Spain
The results of an experimental study is studied in which competitiveness and cooperation are combined to improve efficiency in team production are reported. Teams often suffer from a free-rider problem with respect to individual contributions. Recent literature of experimental economics shows that putting teams into competition with each other can mitigate this problem. However, the question about how the composition of the teams influences on cooperative behaviour, and thereby overall efficiency, remains largely unexplored. The main purpose is to compare team performance between heterogeneous teams (randomly formed) and homogeneous teams (similar preferences) with and without competition between them. As decision framework, we use a public goods game with an all-can-win competition. Individual contribution decisions are repeatedly made during ten periods without competition (independent teams) and, afterwards, during other ten periods with competition (dependent teams) under a within-subject design. Consequently, when analyzing the experimental data, we tackle the dependence problem allowing for clustering at the subject level,

at the team level and at the competition level. Parametric and non-parametric methods are used in the statistical analysis. Our experimental results provide support for a significant impact of both between-team competition and team composition on the overall cooperation level.

E1851: **Evolution in the relation between rainfall and fluvial discharge since 1970 in Esva Basin (NW Spain)**
Presenter: **Elena Fernandez**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Jorge Marquinez, Maria Fernandez-Garcia
The river channel in the Esva basin, a coastal catchment of the Cantabrian region (North Iberian Peninsula) with a surface of 464 km2 and maximum high close to 1300 m, has experienced riverbed's incision, channel narrowing, vegetal colonization and loss of active gravels. This trend has been described in other rivers from Spain and Europe, and they could been linked to a general decrease in sediment supply after land-use changes and/or hydrological changes in rainfall and fluvial discharge. In the latter case, a statistical analysis of the relation between rainfall and fluvial discharge since 1970 is carried out. The impact of the recorded precipitations on the fluvial discharge is analysed based on lagged dependent variable models for different scenarios, trying to detect if there are important hydrological changes.

E1863: **Semi-complete data likelihood for efficient state space model fitting**
Presenter: **Agnieszka Borowska**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Ruth King
A novel efficient model-fitting algorithm for state space models is proposed. State space models are an intuitive and flexible class of models, frequently used due to the combination of their natural separation of the different mechanisms acting on the system of interest: the latent underlying system process; and the observation process. This flexibility, however, comes at the price of significantly more complicated fitting of such models to data as the associated likelihood is typically analytically intractable. For the general case a Bayesian data augmentation approach is often employed, where the true unknown states are treated as auxiliary variables and imputed within the MCMC algorithm. However, standard "vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed. The proposed method circumvents the inefficiencies of the previous approaches by combining data augmentation with numerical integration in a Bayesian hybrid approach. This permits standard "vanilla" algorithms to be applied. The proposed semi-complete data augmentation algorithm is applied to different types of problems demonstrating efficiency gains in empirical studies.

CI003 Room Beveridge Hall SIMULATION-BASED METHODS IN ECONOMICS AND FINANCE Chair: David Frazier

C0176: **Derivative-based optimization with a non-smooth simulated criterion**
Presenter: **David Frazier**, Monash University, Australia
Co-authors: Dan Zhu
Indirect inference requires simulating realizations of endogenous variables from the model under study. When the endogenous variables are discontinuous functions of the model parameters, the resulting indirect inference criterion function is discontinuous and does not permit the use of derivative-based optimization routines. Using a specific class of measure changes, we propose a novel simulation algorithm that alleviates the underlying discontinuities inherent in the indirect inference criterion function, permitting the application of derivative-based optimization routines to estimate the unknown model parameters. Unlike competing approaches, this approach does not rely on kernel smoothing or bandwidth parameters. Several Monte Carlo examples that have featured in the literature on indirect inference with discontinuous outcomes illustrate the approach. These examples demonstrate that this new method gives superior performance over existing alternatives in terms of bias, variance and coverage.

C0175: **Composite indirect inference with application to corporate risks**
Presenter: **Christian Gourieroux**, University of Toronto and CREST, Canada
Co-authors: Alain Monfort
It is frequent to deal with parametric models which are difficult to analyze, due to the large number of data and/or parameters, complicated nonlinearities, or unobservable variables. The aim is to explain how to analyze such models by means of a set of simplified models, called instrumental models, and how to combine these instrumental models in an optimal way. In this respect a bridge between the econometric literature on indirect inference and the statistical literature on composite likelihood is provided. The composite indirect inference principle is illustrated by an application to the analysis of corporate risks.

C0331: **Approximate Bayesian forecasting**
Presenter: **Gael Martin**, Monash University, Australia
Co-authors: David Frazier, Worapree Ole Maneesoonthorn, Brendan McCabe
Approximate Bayesian Computation (ABC) has become an increasingly prominent tool for conducing inference in a range of challenging statistical problems, most notably those characterized by an intractable likelihood function. ABC inference requires only that one can simulate pseudo data from the assumed data generating process underlying the observed data, for given draws of the parameters from the prior. Parameter draws that produce a 'match' between the pseudo and observed data are retained and used to estimate the posterior distribution, with the accuracy of the resultant estimate of the exact (but inaccessible) posterior dependent on the informativeness of the summary statistics used in the matching. We focus on the use of ABC not as a tool for parametric inference, but as a means of generating probabilistic forecasts; or for conducting what we refer to as 'approximate Bayesian forecasting'. The three key issues explored are: i) the loss of forecast accuracy incurred when using an approximate rather than an exact forecast distribution; ii) the role played in approximate Bayesian forecasting by posterior consistency; and iii) the use of forecasting criteria to inform the selection of ABC summaries. A range of time series models, including those in which latent variables, and discrete variables, feature are used to illustrate the methodology.

CO114 Room Bloomsbury RECENT ADVANCE IN NONPARAMETRIC AND SEMIPARAMETRIC ECONOMETRICS Chair: Degui Li

C0412: **Indirect inference for locally stationary models**
Presenter: **Bonsoo Koo**, Monash University, Australia
Co-authors: David Frazier
An indirect inference method is proposed for locally stationary processes. We develop a local indirect inference algorithm and establish the asymptotic properties of the proposed estimator. The convergence rate of the proposed estimator follows the usual semiparametric one due to the employment of the semiparametric kernel estimation procedure. We also validate our methodology via simulation studies. Our local indirect inference method is applied to locally stationary moving average models and locally stationary multiplicative stochastic volatility models.

C0409: **Nonparametric homogeneity pursuit in functional-coefficient models**
Presenter: **Lingling Wei**, University of York, United Kingdom
The homogeneity of coefficient functions in nonlinear models with functional coefficients is explored, and the semiparametric modelling structure is identified. With an initial kernel estimate of each coefficient function, we combine the classic hierarchical clustering method and a generalised

version of the information criterion to estimate the number of clusters each of which has the common functional coefficient and determine the indices within each cluster. To specify the semi-varying coefficient modelling framework, we further introduce a penalised local least squares method to determine zero coefficient, non-zero constant coefficients and functional coefficients varying with the index variable. Through the nonparametric cluster analysis and the penalised approach, the number of unknown parametric and nonparametric components in the models can be substantially reduced and the aim of dimension reduction can be achieved. Under some regularity conditions, we establish the asymptotic properties for the proposed methods such as consistency of the homogeneity pursuit and sparsity. Some numerical studies including simulation and an empirical application are given to examine the finite-sample performance of our methods.

C0624: **Higher order properties of estimators in nonparametric moment conditions models with weakly dependent data**
Presenter: **Francesco Bravo**, University of York, United Kingdom

Localized versions of Generalized Method of Moments (GMM) and Generalized Empirical Likelihood (GEL) are considered that can be used for estimation in nonparametric moment conditions models with weakly dependent data. The local linear estimator is used to estimate the unknown parameters and shows that the local empirical likelihood estimator has the smallest second order bias within the class of local efficient GMM and GEL estimators. Bias corrected local GMM/GEL estimators are proposed that are second order accurate. The results are illustrated analytically and numerically with two examples: a nonparametric generalized linear model and a smooth coefficient model where some covariates can be correlated with the unobservable errors.

C0833: **Adaptive inference in continuous-time asset pricing factor models**
Presenter: **Shin Kanaya**, University of Aarhus, Denmark
Co-authors: Dennis Kristensen, Yang Zu

Using non- and semiparametric techniques, we develop a methodology for estimating and testing continuous-time multi-factor asset pricing models that may potentially contain non-stationary components. The proposed procedures are simple to implement and only involve computing (generalized) least-squares estimators and associated (robust) standard errors. We analyze the asymptotic properties of the estimators and test statistics and show that standard critical values can be employed. A simulation study compare our proposal with existing methods in the literature.

C0755: **Estimating spot variance in possibly explosive AR(1) model: Application to testing bubble with heteroscedastic data**
Presenter: **Yang Zu**, University of Nottingham, United Kingdom
Co-authors: Dave Harvey, Steve Leybourne

Heteroscedasticity, or time-varying volatility/variance, exists widely in economic time series data. Nonparametric estimation is considered for the spot variance function in a varying-coefficient, heteroscedastic AR(1) model, where the AR coefficient of the model dictates a potential change of the model from an unit root regime to an explosive regime. We propose a two-step method to estimate the variance function in the above model. We show the pointwise consistency and the asymptotic normality of the spot variance estimator. We also derive the uniform consistency of the estimator, which is further used in our application in testing explosive bubbles.

CO419 Room Chancellor's Hall INFERENCE FOR HIGH-DIMENSIONAL ECONOMETRIC TIME SERIES Chair: Stephan Smeekes

C0489: **Inference in high-dimensional linear regression models**
Presenter: **Tom Boot**, University of Groningen, Netherlands
Co-authors: Didier Nibbering

An asymptotically unbiased estimator is introduced for the full high-dimensional parameter vector in linear regression models where the number of variables exceeds the number of available observations. The estimator is accompanied by a closed-form expression for the covariance matrix of the estimates that is free of tuning parameters. This enables the construction of confidence intervals that are valid uniformly over the parameter vector. Estimates are obtained by using a scaled Moore-Penrose pseudoinverse as an approximate inverse of the singular empirical covariance matrix of the regressors. The approximation induces a bias, which is then corrected for using the lasso. Regularization of the pseudoinverse is shown to yield narrower confidence intervals under a suitable choice of the regularization parameter. The methods are illustrated in Monte Carlo experiments and in an empirical example where gross domestic product is explained by a large number of macroeconomic and financial indicators.

C0711: **Penalized quasi-maximum likelihood estimation in extended constant conditional correlation GARCH models**
Presenter: **Marco Avarucci**, University of Glasgow, United Kingdom
Co-authors: Stephan Smeekes

In recent years, multivariate GARCH models have become important tools in empirical finance. For instance, asset pricing and risk management crucially depend on the conditional covariance structure of the assets of a portfolio. Modeling volatility spillovers is also crucial to understand the markets connectedness and contagion. The extended conditional correlation (ECCC) GARCH is attractive for its tractability and ease of interpretation; In particular, conditions for the positive definiteness of the conditional variance and for the existence of strictly stationary solutions are simple and explicit However, when the number of assets is large, the quasi-maximum-likelihood (QML) method - arguably the most popular estimation method in the GARCH setting - can be difficult to apply. To handle this kind of problem, regularization techniques are applied. The asymptotic properties of the penalized QMLE are investigated. Special care has to be taken as the zero parameters lie on the boundary of the parameter space, which affects the asymptotic properties of the estimators.

C0555: **Real-time estimation of unemployment with dynamic factor and time-varying state space models**
Presenter: **Caterina Schiavoni**, Maastricht University, Netherlands
Co-authors: Stephan Smeekes, Jan van den Brakel, Franz Palm

Estimation of unobserved components is considered in high-dimensional state space models using a dynamic factor approach. Our method allows for variables to be observed at different frequencies and updates the estimation when new information becomes available. In addition, we account for potential time variation in the parameters of the model. We apply the methodology to unemployment estimation as done by Statistics Netherlands, who uses a multivariate state space model to produce monthly figures for the unemployed labour force using series observed with the Labour Force Survey (LFS). We extend the model by including auxiliary series about job search behaviour from Google Trends and claimant counts, partially observed at higher frequencies. Our factor model allows for nowcasting the variable of interest, providing unemployment estimates in real time before LFS data become available. In addition our method accounts for time-varying correlations between the LFS and auxiliary series.

C1052: **A nonlinear panel structural VARs approach to state dependent fiscal multipliers**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Lenard Lieb

Nonlinearities in dynamic fiscal policy mutlipliers are investigated by using a novel structural panel vector autoregression framework. Nonlinearities are often modeled in the literature through state-dependent linear models that are essentially threshold VARs, where the threshold is linked piecewise to the business cycle. We argue that sycg piecewise linear models may only poorly capture the interaction with the business cycle if nonlinearities are of a more complex nature. Our approach can be seen as a generalization of threshold-type models which allows for a continuum of states.

By exploiting the cross-sectional heterogeneity in the dynamics of the panel we are able to effectively estimate the underlying form of nonlinear dependence of the dynamic responses of the economy to fiscal policy events.

C0465: **Structured regularization of high-dimensional panel vector autoregressions**
Presenter: **Stephan Smeekes**, Maastricht University, Netherlands
Co-authors: Lenard Lieb, Mathias Staudigl
Inference is investigated on large panel vector autoregressions using regularization techniques. Rather than relying on standard sparsity assumptions, our regularization makes use of structured sparsity that exists naturally by similarity of the dynamics within and between groups in the panel. We explicitly allow for spillover effects between groups as our estimation method endogenously provides a measure of the strength of the (dynamic) interconnectedness between the groups. We develop efficient algorithms for estimation and explore the consistency, oracle and inferential properties of our estimator for plausible economic models, while finite sample properties are investigated in a simulation study. Finally, the generality and interpretability of our method is illustrated by an application to the analysis of fiscal policy in a multi-country macroeconomic panel dataset.

CO077 Room Court CENTRAL BANK FORECASTING I Chair: Knut Are Aastveit

C0210: **Forecasting industrial production in Germany**
Presenter: **Sercan Eraslan**, Deutsche Bundesbank, Germany
Co-authors: Klemens Hauzenberger
A parsimonious forecast model is introduced for industrial production in Germany. Our framework is based on the combination of point and density forecasts obtained from a pool with a relatively low number of models capturing, in particular, the long-run dynamics of industrial production. The model pool consists of three model classes: vector AR and error correction as well as multicointegration models. While standard vector error correction models are based on the relation between industrial production and new orders received, the multicointegration framework allows for a second long-run relation in the model: the relation between industrial production and inventories and/or stock of orders. Moreover, we augment all model classes with survey-based indicators specifically targeted at the industrial sector. We consider both linear and threshold-type nonlinear specifications, and estimate the models with Bayesian and frequentist methods. At a final step we combine the point and density forecasts in various ways. Our first results based on a real-time forecast evaluation between 2006 and 2016 points out that the parsimonious forecast routine with around ten models can reduce the root mean squared forecast errors by up to 30% over a 12-month forecast horizon against the random-walk benchmark.

C0411: **Forecasting with VARs with time-variation in the mean**
Presenter: **Andries van Vlodrop**, VU Amsterdam, Netherlands
Co-authors: Marta Banbura
A vector autoregressive model allowing for time variation in the mean and the variance is proposed. The unobserved time-varying mean is assumed to follow a random walk and we also link it to long-run Consensus forecasts, similar in spirit to so called democratic priors. The changes in variance are modelled via stochastic volatility. We propose a Bayesian methodology for inference in the model. The proposed Gibbs sampler allows the researcher to use a large cross-sectional dimension in a feasible amount of computational time. This is in contrast to standard time-varying-parameter VARs, which typically are not used beyond a cross-sectional dimension of three or four variables. Furthermore, standard simulation smoothing methods can be used to perform inference on the latent process of the local mean. The slowly changing mean can account for a number of secular developments such as changing inflation expectations, slowing productivity growth or demographics. We compare the forecasting performance of our model to time invariant VARs with Minnesota and democratic priors and standard time-varying-parameter VARs for several countries. The model performs well against the alternatives. In particular, incorporating survey information through our local mean approach improves the long run forecasting performance of VAR models.

C0611: **On core inflation indicators: Evidence from the European Union countries**
Presenter: **Aleksandra Halka**, Narodowy Bank Polski, Poland
Co-authors: Grzegorz Szafranski
The aim is to decompose the consumer inflation in EU economies into permanent and transitory component to check if selected core inflation measures the underlying trend. In the decomposition we follow a recent method. Second, to address the question of the core inflation role as a demand indicator we construct regressions motivated by the Philips curve. Third, we check the forecasting performance of the selected measure using a previous method. Finally, to summarize the results we propose a synthetic measure Core Inflation Score which evaluates different measures in a simple and comprehensive way. The main results are: Although the energy and unprocessed food prices are volatile they also include part of the permanent component of the inflation. Therefore, the measure that excludes these sub-indices (XUFE) is not the best approximation of the underlying inflation. Despite that the most of the economists argues that the core inflation measures should reflect the trend inflation we also find a positive correlation between XUFE and output gap for some countries. Moreover, for higher number of countries XUFE is approximation of the demand pressure than of the trend inflation. In forecasting future developments of the HICP index in the short term the permanent HICP component may be a better and less biased predictor of the future inflation. At the same time for the longer horizons the permanent XUFE component is better.

C1116: **Financial nowcasts and their usefulness in macroeconomic forecasting**
Presenter: **Edward Knotek**, Federal Reserve Bank of Cleveland, United States
Co-authors: Saeed Zaman
The usefulness of financial nowcasts in making conditional forecasts of macroeconomic variables with quarterly Bayesian vector autoregressions (BVARs) is considered. For nowcasting quarterly financial variables' values, we find the average of the available daily data and a daily random walk forecast to complete the quarter typically outperforms other nowcasting approaches. Using real-time data, we find gains in out-of-sample forecast accuracy from the inclusion of financial nowcasts relative to unconditional forecasts, with further gains from incorporating nowcasts of macroeconomic variables. Conditional forecasts from quarterly BVARs augmented with financial nowcasts rival the forecast accuracy of dynamic factor and mixed-data sampling (MIDAS) models.

C0949: **Residential construction activity and recession predictability**
Presenter: **Andre Anundsen**, Norges Bank, Norway
Co-authors: Knut Are Aastveit
The aim is to assess the importance of residential investments in predicting economic recessions for an unbalanced panel of 12 OECD countries over the period 1960Q1-2014Q4. Our approach is to estimate various probit models with different leading indicators, commonly used in the literature, and evaluate their relative prediction accuracy using the receiver operating characteristic (ROC) curve. As a main result, we document that residential investments contain information useful in predicting recessions both in-sample and out-of-sample. This result is robust to adding typical leading indicators, such as the term spread, stock prices, consumer confidence surveys and oil prices. It is shown that residential investments are particularly useful in predicting recessions for countries with high home-ownership rates. Finally, in a separate exercise for the US economy,

we show that the predictive ability of residential investments is robust to employing real-time data.

CO524 Room G11 BEHAVIOURAL AND EMOTIONAL FINANCE: THEORY AND EVIDENCE Chair: Richard Fairchild

C0240: **Age differences in the effects of metacognition on financial decision-making**
Presenter: **Chiara Scarampi**, University of Bath, United Kingdom
Co-authors: Richard Fairchild, Alberto Palermo, Neal Hinvest
Due to a decline in cognitive function, decision-making abilities can be compromised with advancing age. This is a significant issue given the increasing pressure to take control over financial and personal well-being in old age. A large body of literature has explored the relationship between age and financial choices and the underlying emotional and cognitive processes. However, there are still some gaps in discerning other variables such as metacognition which interact with cognitive and emotional processes when making decisions. The age-related effects that metacognition may have on financial choice behaviour are investigated. The main findings show that young adults have significantly higher cognitive ability, whereas older individuals provide more accurate metacognitive judgements. Furthermore, it emerges from the analysis that some metacognitive components buffer individuals against cognitive decline, providing an alternative route to sound financial decisions. Together, these findings emphasise the importance of studying metacognition in the context of financial decision-making. High metacognitive abilities can enable individuals, and particularly older adults, to reflect on their abilities, on task requirements and on the most appropriate strategies to solve the situation at hand, enhancing in turn competences in making appropriate decisions.

C0253: **A revisit on government ownership: Evidence from China after the secondary privatisation**
Presenter: **Yan Zeng**, Newcastle University Business School, United Kingdom
The role of Government ownership on firm performance in Chinese listed companies before and after the secondary privatisation is discussed. The traditional debate on the helping or grabbing hand of Government fails to capture the dynamic nature of Governments behaviours. We propose a goal alignment hypothesis to reconcile these two paradoxical views and argue that Governments role varies with its strategic goals and policies over time. The main findings manifest that Chinese Government exerts its power in firms corresponding to its long-term principle as well as the time-varying operational polices. Furthermore we find that Government intervention largely props up firms after the secondary privatisation in the light of tremendously decreased ownership and improved governance, which provides a benchmark for politicians and business owners in the leading market economies to draw on, which is especially critical after the massive Government interference in these countries since 2008. In addition, a dynamic system GMM methods is applied to handle potential endogeneity problems so as to maximise the reliability of the results.

C1274: **Analysing financial herding through network analysis**
Presenter: **Muhamed Alsharman**, Bath University, United Kingdom
Co-authors: Richard Fairchild, Neal Hinvest
Emotional Finance introduces the notion that financial markets may be driven by the co-existence of fully-rational and emotional investors, driven by phantasy. The analysis of emotional finance is informed with reference to a Freudian psychoanalytical framework. We add to the existing information cascade and herding research by developing an emotional finance model that examines the effects of phantasy investors on the decisions of rational investors under dynamic pricing. We consider a financial market for a risky asset in which the traders emotions develop overtime based on how they perform. We propose an elementary agent-based asset pricing model consisting of three trader types; fundamental traders, emotional traders, and semi-emotional traders. The model comprises two features: 1) an emotional herding mechanism based on susceptible-infected susceptible (SIS) model 2) wealth price herding based on wealth preferential attachment. Combining analytical and simulation methods, the interaction between these elements is studied in a 4-phase plane of the price movement: 1) prices resembling a bull market 2) prices resembling a bear market 3) U-shaped pricing trends, and 4) n shaped pricing trends. Finally, we compare our approach with a traditional information cascade/herding model incorporating phantasy investors

C0964: **Information cascades, herding and emotional investors in an IPO: Rational decision-making distorted by phantasy**
Presenter: **Richard Fairchild**, University of Bath, United Kingdom
The aim is to add to existing Information Cascade Models and Emotional Finance by developing a herding model that examines the effects of phantastic investors on the decisions of rational investors under dynamic pricing in an IPO scenario. The presence of irrational, phantastic herding in an IPO scenario leads to price inflation caused by irrationality. This reduces the likelihood of rational investors taking part in a new issue. Positive cascades by rational investors are completely eliminated when emotional investors view the new issue as infallible, and in some cases, negative abstaining cascades happen with certainty, regardless of previous rational investors decisions. When emotional investors are overcome by hatred towards the new issue, rational investors are now encouraged to invest in a new issue and, in some cases, will ignore their private information in order to conform with previous rational investors decisions. Thus, emotional investors cause the market to be stressed as they all herd, initially by investing and then abstaining. We find that herding in times of market stress by emotional investors can lead to rational herding, but in the opposite direction.

C0997: **Culture and executive compensation**
Presenter: **Wissam Abdallah**, Lebanese American University, Lebanon
Co-authors: Yang Zhao
Th purpose is to examine how culture impacts executive compensation. Using a sample of 5,545 firms from 16 countries and employing GLOBE cultural dimensions, we find that culture does matter in executive compensation. In particular, we find that in assertive societies CEO demands high compensation and that compensation structure is linked to firm performance. We also observed that in assertive cultures, there is larger compensation disparity among top management team. Moreover, we find that executive compensation is higher and less related to performance in collectivistic societies. We also find higher compensation disparity in the boardroom in collectivistic culture. In addition, we find that managers in countries that are characterised with high uncertainty avoidance have less equity-based compensation. Furthermore, we find that in countries with larger power distance there is more disparity in the top management group. We also investigate the impact of culture on gender pay gap. Our results show that larger proportion of discriminatory compensation gap exists in societies where male and female are more unequal. Lastly, we find that in future oriented societies managers are rewarded with more equity-based compensation, and in societies that are performance oriented, executive compensation is more likely to be linked with performance.

CO304 Room G4 UNCERTAINTY AND CENTRAL BANKS Chair: Svetlana Makarova

C0950: **Measuring the effect of expected inflation uncertainty on the UK economy**
Presenter: **Carlos Diaz**, University of Leicester, United Kingdom
A new measure of ex ante inflation uncertainty based on the revisions of the density forecasts produced by the Bank of England is presented. This measure is interpreted as the expected effect the Bank anticipates that information released between two consecutive revisions (quarters) will have on future inflation uncertainty. Therefore, the measure will be useful to identify uncertainty shocks. The effect of these shocks on a panel of macroeconomic and financial variables is studied.

C0275: **On Fed watchers' eyes: Hawks, doves and monetary Policy**
Presenter: **Klodiana Istrefi**, Banque de France, France

A novel measure of perceived policy preferences for the FOMC is introduced, which reflects the perceptions of market participants as expressed in newspaper articles, financial media outlets and business reports of Fed watchers in the US. Following the language of financial markets, we categorize the information on the preferences of FOMC members in two leanings, hawk and dove, which represent the weights that market participants believe each member assigns to one of the dual objective of the Federal Reserve. Investigating the period 1960-2015, we find substantial variation in the perceived Hawk/Dove scale of the FOMC. Overall, markets have perceived a hawkish FOMC, especially for the most part of Arthur Burns, Paul Volcker and Alan Greenspan's years. Furthermore, a clear dovish bias of the FOMC is perceived during the last years of Fed chairmanship of Martin (second part of the 1960s) and during the years of Ben Bernanke and Janet Yellen. Reserve Bank Presidents are systematically perceived as hawkish and to have more persistent preferences while Board Governors are perceived as swinging more often between types. Overall, these results match well with narratives on monetary policy in the US. Moreover, we observe a good match of the measure of perceived preferences with FOMC voting patterns and with existing proxies for policy preferences.

C0485: **Central banks and economic policy uncertainty: Fueling the fire**
Presenter: **Christopher Hartwell**, Kozminski University, Poland

The post-global financial crisis world has been characterized by two distinct but interrelated phenomena: the advent of unconventional monetary policy and a marked upswing in economic policy uncertainty. For the central bank community, it is perhaps self-evident that unconventional monetary policy has been employed in order to deal with uncertainty shocks, which are themselves a form of aggregate demand shock. However, given the macroeconomic effects of policy uncertainty are relatively small unless amplified and that policy uncertainty actually results in a fall in inflation expectations, it would appear that the upswing in policy uncertainty is not an exogenous phenomenon. In fact, it is the assertion of this paper that the surge in policy uncertainty is directly a product of central banks and especially unconventional monetary policy in the post-crisis world. Rather than calming markets, developed country central banks are instead fueling the fire of uncertainty and acting as the amplification mechanism. Using monthly data on monetary policies, economic policy uncertainty, and macroeconomic conditions around the world, we find that increases in policy uncertainty have indeed been highly correlated with various central bank policies and announcements independent of economic fundamentals. Additionally, in a surprising twist, we find that central banks that are more independent seem to be creating the most policy uncertainty.

C0943: **Central bank transparency and inflation uncertainty: Evidence for South Korea**
Presenter: **Seohyun Lee**, Bank of Korea, Korea, South

Crucial challenges to central bank transparency have been emerged due to high uncertainty since Great Recession. For a small open economy, inflation expectation may be affected by both domestic and global uncertainty. Against this backdrop, the dynamic behaviour of economic policy uncertainty, central bank transparency and inflation is examined by VAR model with South Korean data (April 2002-May 2017). The transparency is measured by the degree of discrepancy between the survey inflation expectation and target in the forward-looking perspective. The endogenous variables are: Economic Policy Uncertainty index, stock market index, interest rate, transparency index, inflation, and industrial production. In addition, three exogenous variables are included: oil price, EPU index of two largest trade partners of Korea, the US and China. We found that an increase in domestic uncertainty reduces transparency. One standard deviation shocks to uncertainty reduce transparency persistently until 9 month horizon with maximum effect of 0.089 decrease. Uncertainty shocks affect inflation to rise but significant only at 8 month horizon. The results confirm uncertainty spillovers from the US and China to Korean economy and the importance of China's economic policy uncertainty on Korean inflation. The forecast errors of inflation are constructed as inflation uncertainty. The inflation uncertainty rejects the Gaussianity.

C1061: **Quasi ex-ante inflation forecast uncertainty**
Presenter: **Svetlana Makarova**, University College London, United Kingdom
Co-authors: Wojciech Charemza, Carlos Diaz

It is argued that the ex-post measure of forecast uncertainty developed from the distribution of inflation forecast errors differs from the corresponding ex-ante measure because of the impact of monetary policy decisions. We derive a proxy for inflation uncertainty, called quasi ex-ante forecast uncertainty, which is to some extend free from the effects of monetary policy decisions. This proxy is computed using the parameters of a weighted skew normal distribution fitted to forecast errors. This in turn leads to the development of the measure of the compound strength of monetary policy and the uncertainty ratio, which shows the relative impact of monetary policy on reducing inflation forecast uncertainty. A nonlinear relationship is found between compound strength and the measures of the independence and transparency of central banks for 38 countries. The quasi ex-ante forecast uncertainty is used for computing the inflation forecast term structure for the BRICS countries (Brazil, Russia, India, China and South Africa), the UK and the US. It is concluded that the greatest policy effect in reducing inflation forecast uncertainty is for countries which conduct either long-established and relatively pure inflation targeting policy, like South Africa and the UK, or clandestine inflation targeting, like India and the US. The smallest reduction is for countries like China and Russia that mix inflation targeting with exchange rate stabilisation.

CO721 Room G5 ADVANCED TOPICS IN MACROECONOMETRIC MODELING Chair: Florian Huber

C0366: **Common versus country-specific factors in Euro Area output gap estimation**
Presenter: **Philipp Piribauer**, Austrian Institute of Economic Research (WIFO), Austria
Co-authors: Florian Huber

Optimum currency area (OCA) theory claims that the synchronization of business cycles is a prerequisite of a well-functioning currency union. In the absence of a certain degree of business cycle synchronicity, policy makers can hardly satisfy the needs of all member states simultaneously. Thus, if member states display a pronounced degree of business cycle heterogeneity, i.e. some countries tend to be in business cycle upturns whereas other countries face a situation where current output is markedly below potential output, designing appropriate area-wide monetary policy proves to be challenging. The aim is to study the nature and the co-movement of economic output trajectories in European countries in general, and its contribution to the European integration process in particular. We use a Bayesian dynamic factor modelling approach as a means to exploit the co-movements of European output trajectories. The modelling approach allows to unveil the properties of the employed common factors and trace their relative contribution over time. Specifically, a quantification of the explained volatility structure due to the establishment of the European Single Market can be assessed both in-sample as well as its relative importance in out-of-sample forecasting exercises.

C0840: **Debt regimes and the effectiveness of monetary policy**
Presenter: **Clara De Luigi**, Vienna University of Economics and Business, Austria
Co-authors: Florian Huber

A medium-scale non-linear model of the US economy is developed. The proposed model, a threshold vector autoregression with stochastic volatility, assumes that changes in government debt-to-GDP ratios drive the transition between regimes, capturing low and high debt regimes. The introduction of hierarchical priors enables us to flexibly shrink the empirical model towards the moments implied by a dynamic stochastic general equilibrium model. In the empirical application, we analyze regime-specific monetary policy shocks. Our findings indicate that the effect of monetary policy is less pronounced in 'high' debt regimes, pointing towards differences in the underlying monetary policy transmission

mechanisms. Forecast error variance decompositions enable us to shed further light on the relative importance of monetary policy shocks within different debt regimes in terms of explaining the variance of forecast errors.

C1193: **Dynamics of the trade balance: In search of the J-curve using a structural gravity approach**
Presenter: **Aurelien Fichet de Clairfontaine**, Vienna University of Economics and Business, Austria
Co-authors: Harald Badinger
In theory, the depreciation of a country's currency is supposed to lead to a negative short-run effect on the trade balance due to increased import prices, and to positive medium-run effects, triggered by subsequent responses to changes in relative prices of imports and exports (i.e., the quantity effect). A quantity effect that is large enough to offset the negative short-run price effects results in a (net) improvement of the trade balance after a depreciation: this relationship is dubbed as J-curve effect. A structural gravity approach is used, specifying currency movements as component of trade costs, in order to derive an empirical trade balance model, which incorporates multilateral resistance terms and accounts for the cross-country variation in the magnitude of the exchange rate pass-through into import (export) price. The model then is estimated using monthly bilateral trade flows between 39 OECD countries, disaggregated into ninety-seven commodity sectors, over the period 2010M1-2015M12. The results support the existence of a pooled J-curve across countries and commodity sectors; at the same time they point to considerable heterogeneity in the effects across countries and industries below the surface of aggregate data.

C0831: **Estimating global crop price elasticities: A VAR approach**
Presenter: **Tamas Krisztin**, IIASA, Austria
Econometric models have been used classically to estimate price elasticities of crops. The importance of taking both the supply and the demand side into consideration when estimating elasticities has been recognized in recent literature. However, only few studies show worldwide coverage and most contribution using differing methods or data sources, thus impeding the comparability of the estimates. Moreover, as we show in a literature review, the range of price elasticities cited in literature for different crop types can vary strongly. Thus, we propose a unified estimation framework for crop price elasticities, based on publicly available time series from FAO. Our approach relies on state of the art modeling techniques, using a factor augmented Bayesian vector autoregression. Through this, our model can take into account a wide range of macroeconomic indicators. Elasticity estimates are presented for 30 global regions and 17 crop types. Moreover, we show that global economic equity indices have a considerable influence on crop prices.

C0640: **Equilibrium credit growth in the euro area: A non-linear model approach**
Presenter: **Thomas Zoerner**, Vienna University of Economics and Business, Austria
A non-linear macroeconometric model is proposed to estimate equilibrium credit growth rates in the Euro Area. A threshold vector error correction model (TVECM) that postulates credit growth fluctuation around some long-run equilibrium value determined by a set of macroeconomic fundamentals. The deviation of credit growth from its fundamental value serves as a threshold variable that determines the transition between three regimes. Our modelling approach allows for estimation of the equilibrium credit growth rate and, in addition, provides insights on whether credit markets appear to be far from their equilibrium values. A flexible Bayesian approach permits fully probabilistic inference and provides regime probabilities that the Euro Area credit market is over or undervalued at a given point in time. Moreover, we investigate the impact of monetary policy on credit growth and its equilibrium value by means of structural impulse response analysis. The results suggest that monetary policy exerts a powerful impact on the wider macroeconomy and is capable of driving credit growth back towards its fair value implied by fundamentals.

CO398 Room Gordon DYNAMICS OF RISK PREMIA Chair: Elise Gourier

C0297: **Time-varying risk premia in large international equity markets**
Presenter: **Hugues Langlois**, HEC Paris, France
The aim is to estimate international no-arbitrage factor models with time-varying factor exposures and risk premia at the individual stock level using a large unbalanced panel of 58,674 stocks in 46 countries over the 1985-2017 period. Multi-factor models with regional and country-specific factors perform well. Factor risk premia vary over time and across countries and are more volatile in emerging markets. The country-specific risk factor premia are important in emerging markets and to a lesser extent in developed markets. Both the four- and the five-factor models capture the factor structure in U.S.-denominated international stock returns.

C0517: **Expected term structures**
Presenter: **Ilaria Piatti**, University of Oxford, United Kingdom
Co-authors: Andrea Buraschi, Paul Whelan
The properties of bond risk premia in the cross-section of subjective expectations are studied. We exploit an extensive dataset of yield curve forecasts from financial institutions and document a number of novel findings. First, the relation between subjective expectations and future realizations is positive, and this result holds for the entire cross-section of beliefs. Second, when predicting short term interest rates, primary dealers display superior forecasting ability when compared to non-primary dealers. When predicting long term rates, however, primary dealers have no information advantage. This suggests that a key source of variation in long-term bonds are risk premia and not short-term rate variation. Fourth, we show that consensus beliefs are not a sufficient statistics to describe the cross-section of beliefs. We build an aggregate measure of bond risk premia based on the beliefs of the most accurate agents, consistent with equilibrium models with disagreement and with Friedmans market selection hypothesis. Finally, we use these ex-ante subjective beliefs to evaluate several reduced-form and structural models. We find support for heterogeneous beliefs models and also uncover a number of statistically significant relationships in favour of alternative rational expectations models once the effect of heterogeneous beliefs is taken into account.

C0941: **Asset prices and portfolio choice with learning from experience**
Presenter: **Christian Heyerdahl-Larsen**, London Business School, United Kingdom
Co-authors: Paul Ehling, Alessandro Graniero
Asset prices and portfolio choice with overlapping generations are studied, where the young disregard history to learn from own experience. Disregarding history implies less precise estimates of output growth, which in equilibrium leads the young to increase their investment in risky assets after positive returns, that is, they act as trend chasers. In equilibrium, the risk premium decreases after a positive shock and, therefore, trend chasing young agents lose wealth relative to old agents who behave as contrarians. The learning from experience based bias implies a welfare cost that is larger than that of output fluctuations. Consistent with findings from survey data, the average belief about the risk premium in the economy relates negatively to future excess returns and is smoother than the true risk premium.

C0752: **Pricey puts and return predictability**
Presenter: **Alex Kontoghiorghes**, Queen Mary University of London, United Kingdom
Significant predictability for the S&P 500 using information extracted from the monthly empirical pricing kernel across a recent ten year sample is documented. These relations persist after controlling for the variance risk premium, risk neutral skewness and kurtosis and the other common behavioural and accounting measures which are used in the predictability literature. The pricing kernel's slope is seen to be highly correlated to

business cycle variables, offering suggestions to where the predictability stems from. Cumulative prospect theory preferences is the advocated driver after time varying risk aversion, sentiment measures and divergences in opinion are ruled out.

C0626: **A cross-sectional analysis of the variance risk premium**
Presenter: **Romeo Tedongap**, ESSEC Business School Paris-Singapore, France
Co-authors: Bruno Feunou, Ricardo Lopez Aliouchkin, Lai Xu
Using a large cross-section of equity options and returns, the total individual firm variance risk premium (VRP) is decomposed into a good and bad component. These two components reflect compensation for upward and downward risk, respectively. We find that firms with a high bad variance risk premium (VRPb) have extremely high average returns. Specifically, going long a portfolio consisting of firms with high VRPb and going short a portfolio of firms with low VRPb yields annual expected returns of 15.7%. This result remains significant in double portfolio sorts where we control for individual firms' skewness and exposure to market skewness. VRPb also helps to explain the cross-section of expected stock returns beyond traditional asset pricing factors, firm characteristics and semi-variances. Furthermore, VRPb is important economically, with a two-standard-deviation increase in VRPb associated with a 19.0% rise in expected annual stock returns. We develop an equilibrium-based asset pricing model that captures these facts.

CO376 Room CLO 102 SMALL-SAMPLE ASYMPTOTICS Chair: Benjamin Holcblat

C0270: **Saddlepoint approximations for short and long memory time series: A frequency domain approach**
Presenter: **Elvezio Ronchetti**, University of Geneva, Switzerland
Co-authors: Davide La Vecchia
Saddlepoint techniques provide accurate, higher order, small sample approximations to the distribution of estimators and test statistics. Except for a few simple models, these approximations are not available in the framework of stationary time series. We contribute to fill this gap by developing new saddlepoint approximations for frequency domain statistics. Under short or long range serial dependence, for Gaussian and non Gaussian processes, we show how to derive and implement our saddlepoint techniques (density and tail areas approximations and tests in the presence of nuisance parameters) for two relevant classes of statistics: ratio statistics and Whittle's estimator. Extensive Monte Carlo experiments illustrate the theory for widely-applied time series models, comparing our new approximations to the ones obtained by first order asymptotic theory and the frequency domain bootstrap. The numerical exercises for Whittle's estimator show that our approximations yield accuracy's improvements, while preserving analytical tractability. Finally, a real data example about the European Central Bank assets dynamics is shown.

C0407: **The empirical saddlepoint estimator**
Presenter: **Fallaw Sowell**, Carnegie Mellon University, United States
Co-authors: Benjamin Holcblat
Previous studies have shown that existing moment-based estimation approaches have poor small-sample performance in some applications. We propose an alternative that is based on the ESP (empirical saddlepoint) approximation of the solutions to the empirical moment conditions. Saddlepoint approximations are known to perform well in small sample. The novel estimator proposed, which we call the ESP estimator, is the mode of the ESP approximation. We show that it is consistent and asymptotically normal, and we study its higher-order bias. We propose novel test statistics based on the ESP estimator. Finally, we also investigate the finite-sample properties of the ESP estimator and related test statistics through Monte-Carlo simulations.

C0402: **Higher order saddlepoint approximations**
Presenter: **Raymond Kan**, University of Toronto, Canada
Co-authors: Simon Broda
Higher order saddlepoint approximations are presented for the density and cumulative distribution function of univariate and multivariate random variables. While the results of higher order saddlepoint approximations for the univariate case are generally known, the corresponding results for the multivariate case are largely missing in the literature. In addition, existing results on multivariate saddlepoint approximations are quite cumbersome to use. As a result, we see very few applications that make use of saddlepoint approximations for the multivariate case. Overcoming this problem, we present simple recursive algorithms that facilitate the computation of higher order terms for the saddlepoint approximation. A number of numerical examples are given to illustrate the potential benefit of using higher order saddlepoint approximations.

C0432: **The second-order bias and MSE of quantile estimators**
Presenter: **Aman Ullah**, University of California Riverside, United States
Co-authors: Tae-Hwy Lee, He Wang
Analytical results are developed on the finite sample properties of quantile estimators. We expand previous results on the second-order bias and MSE of quantile estimators with i.i.d. samples. We discover that for both unconditional and conditional quantile regressions, the median is unbiased for a symmetric unconditional and conditional distribution, and the bias of the other quantiles is larger at the tails of any unconditional and conditional distributions. We point out that the second-order bias will vanish as the sample size increases. The Monte Carlo simulations indicate that the second-order bias corrected estimator has better behavior than the uncorrected ones. An application of the impact of schooling, experience, and tenure on earnings are illustrated in quantile estimation. We find larger bias at the extreme low and high earning quantiles. With our second-order bias correction, the results of the application show the improvement of quantile forecasting.

C0754: **Approximation methods for the Rice formula, with applications to small sample asymptotics**
Presenter: **Anthony Almudevar**, University of Rochester, United States
The Rice formula was originally derived to model the intensity of level crossings made by a smooth stochastic process $X(t)$. In its multivariate extension, where $X(t)$ defines a smooth multidimensional random mapping, the Rice formula is equivalently the intensity function of the point process of solutions to a random system of equations. In this form, it has found application in a wide variety of problems in applied mathematics, physics and mathematical statistics. Although compact in form, evaluation of the Rice formula has proven to be technically challenging, largely because of the inclusion of a conditional expectation of the absolute determinant of a random matrix. We present a number of general higher order approximation methods targeted to this problem. These methods are demonstrated with a number of applications in nonlinear regression and generalized linear models.

CO494 Room Jessel LATEST DEVELOPMENTS IN RISK MODELING AND FORECASTING Chair: Roxana Halbleib

C0190: **Flexible multivariate Hill estimators**
Presenter: **David Veredas**, Vlerick Business School, Belgium
Co-authors: Yves Dominicy, Matias Heikkila
A multivariate family of Hill estimators was recently proposed for elliptically distributed random vectors. We show that the family can be generalized to a broader class and, more importantly, that ellipticity is not required. Only multivariate regular variation is needed. This flexibility in terms of both the estimator and the underlying distribution is possible because regular variation of a random vector is preserved under well-behaved homogeneous functions and, as a corollary, we obtain consistency of the new class of estimators. A Monte Carlo study is conducted to asses the finite sample properties of our estimators under different asymmetric and heavy tailed distributions.

C0216: **Elicitability: The quest of comparing risk measure estimates in a meaningful way**
Presenter: **Tobias Fissler**, University of Bern, Switzerland
Co-authors: Johanna F Ziegel, Tilmann Gneiting
In statistical decision-theory, it is common practice to compare competing point forecasts for unknown future events in terms of loss functions. That is, a forecaster issuing the forecast x is assigned the *loss* $L(x,y)$ if y materializes. Considering forecasts for a certain statistical functional such as the mean, the median, or a risk measure, it is crucial that the loss function used is incentive compatible in the sense that the correctly specified forecast for the functional is the unique minimizer of the expected loss. If a functional possesses such an incentive compatible loss function, it is called elicitable, opening the way to meaningful forecast comparison, but also to M-estimation and regression. Many functionals such as expectiles or quantiles are elicitable, whereas other important quantities such as the variance or Expected Shortfall fail to be elicitable. Nevertheless, they can be components of elicitable vector-valued functionals, in particular, both mean and variance, as well as Value-at-Risk and Expected Shortfall are *jointly* elicitable. The latter result opens the possibility to *comparative* backtests aiming at model selection rather than model validation which turns out to be beneficial in the context of quantitative risk management.

C0692: **Regression based expected shortfall backtesting**
Presenter: **Timo Dimitriadis**, University of Konstanz, Germany
Co-authors: Sebastian Bayer
A new backtest for the risk measure Expected Shortfall is introduced. This backtest is based on a Mincer-Zarnowitz regression using a joint linear regression technique for the quantile and the Expected Shortfall. Developing accurate backtests for the Expected Shortfall is particularly relevant in light of the recent swap from Value at Risk to Expected Shortfall in the Basel regulatory framework for banks. We compare the empirical performance of our new backtest to existing approaches in terms of its size and power by simulating returns stemming from standard financial time series models such as the GARCH and the autoregressive stochastic volatility model. Our new backtest exhibits better size and (size-adjusted) power properties compared to existing backtesting procedures in the literature. This shows that our backtest is superior in determining whether banks issue correct risk forecasts for their financial products. We apply this backtesting procedure to Expected Shortfall forecasts for the historical S&P500 index return series. Furthermore, we provide an R package for this backtest which is easily applicable for practitioners.

C0736: **Intraday variance-covariance matrix estimation: A point process approach**
Presenter: **Ingmar Nolte**, Lancaster University, United Kingdom
Co-authors: Sandra Nolte, Yifan Li
The point process based volatility estimator provides an important alternative to the popular Realized Variance (RV)-type estimators in estimating the high-frequency volatility. It has been shown that the volatility estimates from the duration or intensity-based model performs at least equally well with the estimates from the RV-type models. Moreover, this type of volatility estimator has several advantages over the RV-type ones, in particular the fully parametric volatility estimation which utilizes data beyond the daily aggregation window. The parametric design also allows for the inclusion of other market microstructure covariates. We propose an estimator of the high-frequency variance-covariance matrix based on a point process approach by extending the univariate intensity-based framework. We show that our estimator provides reliable variance-covariance estimates while inheriting the advantages of the univariate version over the realized approach. In principle, our estimator can be applied to any sets of price series that provide enough observations in each series for intraday covariance estimation. The parametric structure allows intraday covariance inference that takes the autoregressive structure and diurnal pattern into consideration.

C1818: **Modelling realized covariance matrices by means of factor models**
Presenter: **Roxana Halbleib**, University of Konstanz, Germany
Co-authors: Giorgio Calzolari
A latent factor model is proposed to capture the dynamics of daily realized covariance matrix series with forecasting purposes. The long memory in the series is captured by means of aggregating latent factors with short memory, where the factors are extracted from the common dynamics of realized variance and covariance series. Our approach accommodates the positive-definiteness of the variance-covariance matrix forecasts within a very parsimonious framework. For estimation purposes, we implement the quasi maximum likelihood approach applied on the Kitagawa state-space filtering procedure. We provide Monte Carlo evidence on the accuracy of the estimates and real data evidence on the very good performance of the model to forecast variance-covariance matrices one-step and multi-step ahead.

CO112 Room Montague REGIME SWITCHING, FILTERING, AND PORTFOLIO OPTIMIZATION Chair: Leopold Soegner

C0953: **Enhancing trading strategies under regime shifts**
Presenter: **Christina Erlwein-Sayer**, Fraunhofer Institute for Industrial Mathematics ITWM, Germany
Co-authors: Tilman Sayer, Gautam Mitra
A Markovian regime switching model is applied to identify market regimes in equity markets. Hidden regimes are commonly modelled through Hidden Markov Models. To detect non-observable market regimes, we apply filtering techniques to filter these possible market states out of the observation process. We use the detected regime to choose suitable long and short selling ratios within the portfolio for a daily trading strategy, which is based on a Second Order Stochastic Dominance criterion. We analyse major indices, find most probable state sequences through the Viterbi algorithm and find suitable trading strategies for each regime. The performances of the trading strategies are computed using well established static and dynamic measures like Sharpe and Sortino ratios as well as max drawdowns and days to recovery. We find that the consideration of regime shifts improves the performances of our trading strategies.

C0685: **Local alternatives of signal detection tests**
Presenter: **Klaus Poetzelberger**, WU Vienna, Austria
A principal components analysis of local alternatives of certain one-sided tests for the signal detection problem is presented. These tests reject a null hypothesis if a certain diffusion process crosses a specified boundary b. Identifying local alternatives with high power is thus equivalent to identify directions from b with maximal increasing boundary crossing probabilities (BCP). In cases, where the BCP is given in closed form, the

alternatives can be computed analytically. In all other cases only numerical solutions are possible. We discuss how local alternatives with high power depend on the choice of b.

C0680: **Analyzing and testing the triangular arbitrage parity**
Presenter: **Julia Reynolds**, University of Lugano, Switzerland
Co-authors: Leopold Soegner, Martin Wagner, Dominik Wied
New econometric tools are developed that allow for the monitoring of deviations from arbitrage parities. The econometric methodology is applied in order to test for the stability of the Triangular Arbitrage Parity. Economically relevant variables that possibly result in arbitrage parity violations are analyzed to connect our empirical results to the limits to arbitrage literature. In addition, we also check whether (possible) arbitrage parity deviations can be connected to historical events such as the failure of Lehman Bros. or to central bank interventions.

C0967: **Model reduction and filtering for portfolio optimization in hidden Markov models**
Presenter: **Joern Sass**, University of Kaiserslautern, Germany
A regime switching model, where the observation process is a diffusion process whose drift and volatility coefficients jump governed by a continuous-time Markov chain, can explain some of the stylized facts of asset returns. In the special case that the volatility is constant, the underlying Markov chain can no longer be observed and has to be estimated by its filter. Portfolio decisions then depend on this filter and its dynamics. In fact it turns out that optimal portfolio policies and filter equations rely on the same signal to noise matrix. This can be used to reduce the dimension of the model to the dimension of this matrix if it has full rank. The eigenvalues of this matrix then provide a way to decompose the optimal portfolio in investments in mutual funds. In contrast to classical mutual fund theorems in continuous time, their composition is constant over time but the optimal policy is not. We provide convergence and decomposition results for optimization and filtering. Further we analyze the case of signal to noise matrices which are not of full rank and look at extensions to regime switching models and to hidden Markov models with non-constant volatility. We discuss consistency of the corresponding discrete-time and continuous-time models in view of filtering and portfolio optimization.

C0758: **Expert opinions and their approximation for multivariate stock returns with Gaussian drift**
Presenter: **Dorothee Westphal**, University of Kaiserslautern, Germany
Co-authors: Joern Sass, Ralf Wunderlich
A financial market with multivariate stock returns where the drift is an unobservable Ornstein-Uhlenbeck process is investigated. Information is obtained by observing stock returns and unbiased expert opinions. The optimal trading strategy of an investor maximizing expected logarithmic utility of terminal wealth depends on the conditional expectation of the drift given the available information, the filter. We investigate properties of the filters and their conditional covariance matrices. This includes the asymptotic behaviour for an increasing number of expert opinions on a finite time horizon and conditions for convergence on an infinite time horizon with regularly arriving expert opinions. In the situation where the number of expert opinions goes to infinity on a finite time horizon we distinguish between the case where experts have some minimal level of reliability and experts whose uncertainty increases with increasing frequency of information dates. The latter case leads to a diffusion approximation where the limiting diffusion can be interpreted as a continuous-time expert. This approximation for high-frequency experts thus allows to work with a simpler model in which more explicit solutions can be derived. We deduce properties of the value function using its representation as a functional of the conditional covariance matrices.

CO526 Room Senate CO-INTEGRATION, TREND BREAKS, AND MIXED FREQUENCY DATA Chair: Robert Taylor

C1200: **Bootstrapping non-stationary stochastic volatility**
Presenter: **Peter Boswijk**, University of Amsterdam, Netherlands
Co-authors: Giuseppe Cavaliere, Iliyan Georgiev, Anders Rahbek
Recent research has shown that the wild bootstrap delivers consistent inference in time-series models with persistent changes in the unconditional error variance. Consistency means that bootstrap p-values are asymptotically uniformly distributed under the null hypothesis. The question whether this consistency result can be extended to models with non-stationary stochastic volatility is addressed. This includes near-integrated exogenous volatility processes, as well as near-integrated GARCH processes, where the conditional variance has a diffusion limit. We show that the conventional approach, based on weak convergence in probability of the bootstrap test statistic, fails to deliver the required result. Instead, we use the concept of weak convergence in distribution. Using this concept, we develop conditions for consistency of the wild bootstrap for testing problems with non-pivotal test statistics. Examples are the sample average of a martingale difference sequence, and unit root test statistics with martingale difference errors, both in the presence of non-stationary stochastic volatility. An important condition for wild bootstrap validity is the absence of statistical leverage effects, i.e., correlation between the error process and its conditional variance. The results are illustrated using Monte Carlo simulations and an empirical application.

C1159: **Frequency domain estimation of cointegration vectors with mixed frequency and mixed sample data**
Presenter: **Marcus Chambers**, University of Essex, United Kingdom
The asymptotic properties of band-limited spectral regression estimators of cointegration vectors are derived when mixed frequency data are available. The finite sample performance of the estimators is assessed in a simulation exercise which also considers the properties of a Wald statistic. An optimal sub-system approach that does not require a preliminary, consistent estimator of a spectral density matrix is found to have particularly good finite sample properties.

C0762: **Adaptive information-based methods for determining the co-integration rank in heteroskedastic VAR models**
Presenter: **Luca De Angelis**, University of Bologna, Italy
Co-authors: Robert Taylor, Giuseppe Cavaliere, Peter Boswijk
Standard methods for determining the co-integration rank of vector autoregressive (VAR) systems of variables integrated of order one are affected by the presence of heteroskedasticity with sequential procedures based on Johansen's (pseudo-)likelihood ratio [PLR] test being significantly over-sized in finite samples and even asymptotically. Notable solutions to this problem are the wild bootstrap applied to the PLR test or an information criterion such as BIC. However, although asymptotically valid, these methods may show low power in small samples as they do not exploit the potential efficiency gains provided by the adaptation with respect to the volatility process. Adaptive methods where the covariance matrix is estimated non-parametrically can be particularly useful in the determination of the co-integration rank in VAR models driven by heteroskedastic innovations. Adaptive information criteria-based approaches can also been used to jointly determine the co-integration rank and the VAR lag length. It is in fact well-known that an incorrect selection of the number of lags has a relevant impact on the efficacy of information criteria as well as standard and bootstrap PLR tests in determining the co-integration rank in finite samples. We show that adaptive information criteria are weakly consistent for co-integration rank and lag length determination in the presence of non-stationary unconditional heteroskedasticity, provided the usual conditions on the penalty term hold.

C0836: **Level shift estimation in the presence of non-stationary volatility with an application to the unit root testing problem**
Presenter: **David Harris**, University of Melbourne, Australia

The aim is to investigate the properties of the standard residual sum of squares (RSS) based estimators for the location of a level break in cases where the driving innovations are heteroskedastic, displaying non-stationary volatility (permanent changes in unconditional volatility) and/or conditional heteroskedasticity. Although designed for homoskedastic innovations, the RSS estimator retains its usual rates of consistency under such forms of heteroskedasticity. However, we present simulation evidence which highlights the potential for a very serious decline in the finite sample performance of the RSS estimator, relative to the homoskedastic case, when heteroskedasticity is present. As a consequence, we explore a weighted version of the estimator based around an adaptive estimate of the volatility path of the innovations. The consistency of the weighted estimator is demonstrated and their finite sample behaviour explored. Where the level break is located in a low volatility regime this is shown to deliver very significant improvements.

C1158: **Testing for co-integration rank in VAR models allowing for multiple breaks in trend and variance**
Presenter: **Robert Taylor**, University of Essex, United Kingdom
Co-authors: Giuseppe Cavaliere, David Harris, Simon Price

The problem of testing for the co-integration rank of a VAR process is considered in environments where multiple breaks can occur in the deterministic trend. Unlike existing procedures, we do not assume the break locations or the number of trend breaks to be known. Moreover, we allow the driving shocks to display non-stationary volatility and/or conditional heteroskedasticity, in each case of an unknown form. We use information criteria to select the VAR lag order and the number of trend breaks to fit, using likelihood-based break fraction estimators. Both the information criteria and the break fraction estimators are based around an adaptive (non-parametric) estimator of the variance matrix of the shocks to allow for any non-stationary volatility present. We show that our proposed adaptive information criterion consistently selects the lag length and the correct number of trend breaks in large samples, the locations of which are also consistently estimated. Based on these outcomes, wild bootstrap implementations of Johansen's likelihood ratio tests are then constructed. We show that these deliver asymptotically correctly sized and consistent inference on the co-integration rank regardless of the number of trend breaks present in the data and in the presence of heteroskedasticity.

CO565 Room Woburn MODELING AND ESTIMATING BUSINESS AND FINANCIAL CYCLES Chair: Andrea Silvestrini

C1014: **Real and financial cycles: Estimates using unobserved component models for the Italian economy**
Presenter: **Andrea Silvestrini**, Banca d'Italia, Italy
Co-authors: Guido Bulligan, Lorenzo Burlon, Davide Delle Monache

The empirical features of both the business and financial cycles in Italy are examined. We employ univariate and multivariate trend cycle decompositions based on unobserved component models. Univariate estimates highlight the different cyclical properties (persistence, duration, and amplitude) of real GDP and real credit to the private sector. Multivariate estimates uncover the presence of feedback effects between the real and financial cycles. At the same time, in the most recent period (2015-2016), the multivariate approach highlights a wider output gap than that estimated by the considered univariate models.

C1472: **Semi-structural credit gap estimation**
Presenter: **Peter Welz**, European Central Bank, Germany
Co-authors: Jan Hannes Lang

A theory-based approach is proposed to identifying excessive household credit developments. An equilibrium-relationship for the level of household credit is derived using an overlapping generations model that takes into account the demographic age structure of the economy. In the scope of this framework it is shown that potential GDP, a measure of the equilibrium real interest rate, information about the debt distribution by cohorts and the level of institutional quality are important determinants of the trend household credit stock. Semi-structural household credit gaps are obtained as deviations of the real household credit stock from this fundamental trend level. The resulting model is estimated as an unobserved components system for 12 EU countries using quarterly data for the period 1980 - 2015. Estimates of these credit gaps yield credit cycles that last between 15 to 25 years with amplitudes of around 20%. The estimated credit gaps possess superior early warning properties for financial crises relative to a number of established statistical credit gaps, notably the commonly used Basel total-credit-to-GDP gap and its household credit-to-GDP gap variant. The proposed semi-structural household credit gaps could therefore provide useful information for the formulation of countercyclical macroprudential policy.

E0869: **Business confidence indicators across (similar) surveys**
Presenter: **Tatiana Cesaroni**, Bank of Italy, Italy
Co-authors: Libero Monteforte

Business survey indicators are widely used by policy makers to predict real economy evolution in the short run. The aim is to compare the predictive content of three business survey indicators collected in two different Italian business surveys, namely employment expectations, judgements concerning the general economic situation and credit conditions. More in detail we compare the cyclical properties of these business survey indicators and their ability to provide reliable forecasts with respect to the real economy using both spectral methods and forecasting models (i.e. VAR models).

C1526: **Financial cycles in the euro area: A wavelet analysis**
Presenter: **Michael Scharnagl**, Deutsche Bundesbank, Germany
Co-authors: Martin Mandler

The aim is to study the relationship of loans to non-financial corporations, loans to households, house prices and equity prices between countries or within countries for nine euro area countries. Wavelet analysis allows to account for variations in these relationships both over time and across frequencies. We find evidence of strong co-movements between the growth rates in loans to non-financial corporations over the sample from 1980 to 2015 for all nine countries. For loans to private households the co-movement increases with the start of EMU. The cycle length is in general shorter than claimed by the BIS. Equity prices co-move at various frequencies. The degree of synchronization is in general higher for specific series across the EMU countries than across financial series within specific countries.

C1473: **Financial frictions and business cycles: A wavelet analysis**
Presenter: **Duarte Maia**, University of Minho, Portugal

Using the a model with financial frictions and wavelet analysis, we explore the time-frequency relationship between the credit standards and GDP, debt repurchase, equity payout, hours worked, wages, investment, capital and consumption. We find out that the credit standards have a negative impact on GDP, investment and consumption and a positive impact on debt repurchases, wages and capital. We also find that the model at overall capture the time-frequency features of the empirical time series when we make use of the Wavelet Coherency and Partial Wavelet Coherency.

CO400 Room SH349 COMPUTATIONAL ECONOMETRICS AND MODELLING I Chair: Gareth Peters

C1487: **Which risk factors drive oil futures price curves: Speculation and hedging in the short and long-term**
Presenter: **Matthew Ames**, Institute of Statistical Mathematics, Japan
Co-authors: Guillaume Bagnarosa, Gareth Peters, Pavel Shevchenko, Tomoko Matsui
A consistent estimation framework is developed, building on a well-known two-factor model, to allow for an investigation of the influence of observable factors, such as inventories, production or hedging pressure, on the term structure of crude oil futures prices. Using this novel Hybrid Multi-Factor (HMF) model, we can obtain closed form futures prices under standard risk neutral pricing formulations, and importantly we can incorporate state-space model estimation techniques to consistently and efficiently estimate the models developed. In particular, under the developed class of HMF models and their corresponding estimation framework both the structural features related to the convenience yield and spot price dynamics (or equivalently the long and short term stochastic dynamics) and also the structural parameters that relate to the influence on the spot price of the observed exogenous factors. We can utilize such models to gain significant insight into the futures and spot price dynamics in terms of interpretable observable factors that influence speculators and hedgers heterogeneously, which is not attainable with existing modelling approaches.

C1386: **Bayesian inference for dynamic cointegration models with application to soybean crush spread**
Presenter: **Maciej Marowka**, Imperial College London, United Kingdom
Co-authors: Gareth Peters, Nikolas Kantas, Guillaume Bagnarosa
In crush spread commodity trading strategies, it is a common practice to select portfolio positions based on physical refinery conditions and efficiency in extracting byproducts from crushing raw soybeans to get soyoil and soymeal. The selected portfolio positions are then used to provide a basis for constructing the so called spread series, which is investigated separately using a model with a linear Gaussian structure. We take a statistical approach instead based on forming portfolio positions following from the cointegration vector relationships in the price series. We propose an extension of the standard Cointegrated Vector Autoregressive Model that allows for a hidden linear trend under an error correction representation. The aim is to perform Bayesian estimation of the optimal cointegration vectors jointly with latent trends and to this end we develop an efficient Markov Chain Monte Carlo (MCMC) algorithm. The performance of this method is illustrated using numerical examples with simulated observations. Finally, we use the proposed model and MCMC sampler to perform analysis for soybean crush data. We will find the evidence in favour of the model structure proposed and present empirical justification that cointegration portfolio selection based on physical features of soybean market is sensitive to different roll adjustment methods used in the industry.

C1380: **Spatio-temporal modeling of yield-weather dependence**
Presenter: **Guillaume Bagnarosa**, Rennes School of Business, France
Co-authors: Suikai Gao, Gareth Peters, Matthew Ames, Tomoko Matsui
Weather risk represents today one of the main challenge for the farming businesses and, consequently, for the insurance companies proposing them hedging solutions. However, the lack of data at the farm level, and the ubiquitous moral hazard associated to crop insurance make the pricing of insurance policies quite perilous and requires thus new approaches to solve both problems. We propose to first combine a farm acreage weighted Gaussian Process with a SARIMA dynamic to cope with the weather spatio-temporal joint distribution which is then used to model the distribution at the county level of the crop yields. We apply our new methodology to a large database of Romanian farms and demonstrate that our approach outperforms other methods commonly used in the industry for determining insurance premium rate.

C1647: **Asymptotics for differentiated product demand/supply systems with many markets in the presence of national micro moments**
Presenter: **Yuichiro Kanazawa**, International Christian University, Japan
Co-authors: Yuichiro Kanazawa
Asymptotic properties have been previously derived for an estimator of a supply and demand model in a single market extended with additional moments relating consumer demographics to the characteristics of purchased products as the number of available products increases. We investigate similar asymptotic properties for supply and demand models as researchers observe many more markets and at the same time, national micro moments are available to them. We provide conditions guaranteeing the asymptotic theorems hold for the random coefficient logit model of demand with oligopolistic suppliers. Extensive simulation studies demonstrate significant benefits of the micro moments in estimating the random coefficient logit model.

C1368: **Robust probabilistic PCA and dynamic factor models in mortality**
Presenter: **Gareth Peters**, University College London, United Kingdom
Co-authors: Dorota Toczydlowska, Pavel Shevchenko
A multi factor extension of the family of Lee-Carter stochastic mortality models is developed. We build upon the time, period and cohort stochastic model structure to extend it to include exogenous observable demographic features that can be used as additional factors to improve model fit and forecasting accuracy. We develop a dimension reduction feature extraction framework which a) employs projection based techniques of dimensionality reduction; in doing this we also develop b) a robust feature extraction framework that is amenable to different structures of demographic data; c) we analyse demographic data sets from the patterns of missingness and the impact of such missingness on the feature extraction, and d) introduce a class of multi-factor stochastic mortality models incorporating time, period, cohort and demographic features, which we develop within a Bayesian statespace estimation framework; finally e) we develop an efficient combined Markov chain and filtering framework for sampling the posterior and forecasting.

CC705 Room MAL 152 CONTRIBUTIONS IN COMPUTATIONAL ECONOMETRICS Chair: Ana Escribano

C1699: **Intersectorial default contagion: A multivariate Poisson auto-regression analysis**
Presenter: **Ana Escribano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Mario Maggi
Credit rating default dependences are analyzed in a multi-sectorial framework. Using the Mergent's FISD database, we study the default series in the US over the last two decades, disaggregating defaults by industry-sector groups. During this period, two main waves of default occurred: the implosion of the "dot com" bubble, and the global financial crisis. We estimate a Multivariate Autoregressive Conditional Poisson (MACP) model to the weekly number of defaults occurred in the different sectors of the economy. We discuss the contagion effect between sectors in two ways: the degree of transmission of the probability of default from one sector to another, i.e., the 'infectivity' of the sector, and also the degree of contagion of one sector from another, i.e., the 'vulnerability' of the sector. Our results show some differences between the sectors' relations during the first and the second part of our sample. We add to the analysis some exogenous variables and evaluate their contribution to the goodness of fit.

C1708: **Optimal model selection in binary predictive data science models**
Presenter: **Marta Galvani**, University of Pavia, Italy
Co-authors: Silvia Figini
A novel methodology is presented to assess predictive models for a binary target. One of the main weakness of the criteria proposed in the literature

is not to take the financial costs of a wrong decision into account. The objective is to improve model assessment and selection. The methodological proposal can be of interest for a wide range of applications.We describe how our proposal performs in a real application in credit risk.

C1764: **Joint estimation of parameters of mortgage portfolio and the factor process**
Presenter: **Jaroslav Dufek**, Institute of Information Theory and Automation, Czech Republic
Co-authors: Martin Smid
A factor model for LGD (loss given default) and PD (probability of default) of mortgage portfolio based on KVM approach has been previously proposed. An evolution of factors by a VECM model has also been fitted; however, the parameters of a portfolio are taken as fixed instead of estimation. We propose a technique of a joint estimation of VECM and portfolio parameters in particular MLE function is defined; asymptotic properties are discussed. We propose a technique for joint estimation of the VECM and the portfolio parameters. In particular, MLE function is defined and its asymptotic properties are discussed. Finally, our technique is applied to US market data.

C1465: **Factor models by frequency by multivariate circulant singular spectrum Analysis - MCSSA**
Presenter: **Eva Senra**, Universidad de Alcala, Spain
Co-authors: Pilar Poncela, Juan Bogalo
Circulant Singular Spectrum Analysis (CSSA) allows the extraction of the unobserved components associated with any ex-ante desired frequency in a time series in an automatic way. We first generalize the technique to a multiple setup and automatize it in the same way by the use of circulant matrices applied to the new multivariate trajectory matrix. Second, we extend our proposed methodology to perform factor analysis by frequency. In this way we can jointly extract common factors associated to the trend, to the cycle or to the seasonal component. Finally, we show the application of MCSSA to find common cycles in economic data by sectors or countries.

C1662: **Maximizing discriminatory power of bankruptcy prediction models: Empirical evidence over short and long-term horizons**
Presenter: **Christakis Charalambous**, University of Cyprus, Cyprus
Co-authors: Spiros Martzoukos, Zenon Taoushianis
Acknowledging the economic benefits associated with the development of powerful bankruptcy prediction models, a methodology is presented for maximizing their discriminatory power over short and long-term horizons. For our analysis, we use accounting and market-related information for a sample of U.S. public bankrupt and healthy firms between 1990 and 2015. Results show an improvement in the discriminatory power when we implement our approach as compared with traditional approaches, such as logistic regression models, using short (one and two years) and long-term (five years) forecasting horizons. Most importantly, this improvement in model performance is evident not only in-sample but also when employing three out-of-sample approaches and in several cases is substantial, even when making longer-term forecasts.

CC713 Room MAL 153 CONTRIBUTIONS IN EMPIRICAL FINANCE Chair: Robert Jung

C1611: **A tale of sentiment driven tail events: A dynamic quantile model for asset pricing with sentiment**
Presenter: **Jozef Barunik**, UTIA AV CR vvi, Czech Republic
Co-authors: Cathy Yi-Hsuan Chen, Wolfgang Karl Haerdle
The link between investor sentiment and asset valuation is the subject of considerable debate in the profession. The aim is to abandon the classical asset pricing that relies on expected utility, and introduce a dynamic quantile model for asset pricing, in which the agent maximizes stream of the future quantile utilities instead. Using the model, we empirically investigate if investor sentiment distilled from textual mining analysis can price tails of the return distributions. On the panel of 100 stocks, we document influence of aggregate investor's sentiment on future conditional quantiles of the return distributions. Aggregate sentiment explains cross-section of tails even after controlling for popular factors used in the literature, as well as firm-specific sentiment and volatility.

C1720: **Estimation of market impact cost using high frequency execution and order book data**
Presenter: **Kenta Yamada**, National Institute of Informatics, Japan
Co-authors: Takayuki Mizuno
The aim is to analyze the historical data set of the Tokyo Stock Exchange(TSE) for a 29-month period from August 2014 to December 2016 which includes every transaction and order book snapshot, and found two major relationships: (i) a proportional relationship between the return of the market price and the order imbalance between buying and selling market orders, where this relationship describes market impact, and (ii) an inverse proportional relationship between the market impact and the averaged volumes of limit orders on the order book. In this analysis, we focus on daily and monthly time scale, and we found the market impact depends on the stock and period, however applying the binomial test to the size of the market impact among stocks, it is statistically shown that market impact depends on the stocks. In the analysis of (ii), we show the market impact depends on the order book volume. The results of our analysis imply that when we estimate transaction costs, we need to consider not only a commission fee and spread cost but also a market impact cost, and therefore we also introduce estimation results of transaction costs for a practical application of our studies.

C0237: **The information content of stock market factors**
Presenter: **David McMillan**, University of Stirling, United Kingdom
The information content within several popular stock market factors is examined. Specifically, do they contain independent explanatory power for stock returns and are their movements related to economic variables. Given the explosion in the number of such factors, it is important to consider whether each factor adds to our understanding of markets and the wider economy. We undertake three related exercises, whether a set of popular factors contain independent predictive ability for stock return portfolios, what economic and market variables can explain movements in the factors and whether stock market factors have predictive power for future output growth. The results support the view that while many factors are suggested in the literature, they do not necessarily contain independent information for stock returns. Further, most of these factors do not provide any predictive power for future output growth and thus may not contain any information with regard to economic behaviour although it is apparent is that movement in the factors is driven by considerations of current economic and market risk. Hence, stock market factors reflect existing economic conditions but do not provide a window on future behaviour.

C1160: **Confirming signals are hard to resist: Blessing and curse of information under confirmation bias**
Presenter: **Stefanie Schraeder**, UNSW, Australia
According to empirical evidence, individuals pay more attention to confirming than to contradicting information. In the context of this confirmation bias, we study the effects of additional information on perception correctness – contrasting the competing effects of total signal precision and the possibility to search for the most suitable signal. We provide the testable hypothesis that managers report bad news in a more diffuse signal compared to good news. This, in turn, provides a rationale for the dispersion anomaly: dispersed analysts' earnings forecasts are followed by stock under-performance. Then, we include the confirmation bias in an overlapping generations model with a continuous signal distribution. Several results of more simplified models do not hold any longer. For instance, confirmation bias leads to underreaction instead of overreaction. A momentum effect is accompanied by various time-varying market participation, volatility, trading volume, and market depth effects.

C0238: **Detecting market irrationality using news sentiment and information entropy**
Presenter: **Anqi Liu**, Cardiff University, United Kingdom
Co-authors: Jing Chen, Steve Yang
News sentiment, an investor sentiment proxy, has been widely explored in behavioral finance; yet the linkage between investor sentiment and market irrationality and inefficiency has not been thoroughly examined. We consider the financial market as a bivariate system that consists of news sentiment and market returns. We adopt the concept of transfer entropy to quantify information flow between these two types events and formulate irrationality regime proxies. Testing with an intraday dataset from 2003 to 2014 for the major U.S. markets, we find that the information flow follows a trimodal distribution that clearly distinguishes financial markets into three regimes: the price-driven, transitional and news-driven regimes. We provide evidence to show that the proposed irrationality proxy is positively correlated with three market inefficiency indicators in the current literature; and also we identify a significant cut-off threshold to delineate the market into price-driven and news-driven regimes, showing that news-driven investment decision is a key factor of market inefficiency.

CG245 Room G3 CONTRIBUTIONS IN TESTING AND FORECASTING FINANCIAL TIME SERIES Chair: Laurent Pauwels

C0818: **Justifying conditional confidence intervals using sample splitting**
Presenter: **Alexander Heinemann**, Maastricht University, Netherlands
Co-authors: Eric Beutner, Stephan Smeekes
In order to properly quantify uncertainty around point forecasts and point estimates of objects conditional on the observed data (such as conditional means or variances), parameter estimation uncertainty has to be taken into account. Attempts to incorporate parameter uncertainty are typically based on the unrealistic assumption of observing two independent processes, where one is used for parameter estimation, and the other conditioned upon to obtain forecasts. Such unrealistic foundations raise the question whether these intervals are actually theoretically justified in a realistic setting. This issue is addressed and an asymptotic justification is provided for these type of intervals that does not require such an unrealistic assumption. Our proposed solution is based on a simple sample-splitting approach, which allows us to construct asymptotically valid intervals without relying on the assumption of observing two samples. By showing that our sample-split intervals coincide asymptotically with the standard intervals, we provide a novel, and realistic, justification for confidence intervals of conditional objects, that extends to prediction intervals. The analysis is embedded in the context of Markov chains nesting several important models such as ARMA and GARCH.

C1278: **Optimal combination of forecasts under mean absolute deviation**
Presenter: **Laurent Pauwels**, University of Sydney, Australia
Co-authors: Felix Chan
Theoretical motivations are presented for combining forecasts optimally under the Mean Absolute Deviation (MAD) criterion. While the literature has covered extensively the optimal combination of forecasts under Mean Squared Errors (MSE) theoretically and empirically, it is sparse with respect to MAD. Under normality, the optimization problem for MAD is demonstrated to yield the same solutions as the optimization problem for MSE. The forecast errors are normally distributed as the number of forecasts combined and/or the forecasting sample are increasingly large. Furthermore, a set of conditions for which the simple average is the optimal weight is provided. Simulation studies support the theoretical results and show their relevance when combining a large number of forecasts and/or for long time-series.

C1567: **Inference from the futures: Ranking the noise cancelling accuracy of realized measures**
Presenter: **Giorgio Mirone**, CREATES & Aarhus University, Denmark
The aim is to consider the log-linear relationship between futures contracts and their underlying assets and show that in the classical Brownian semi-martingale (BSM) framework the two series must, by no-arbitrage, have the same integrated variance. We then introduce the concept of noise cancelling and propose a generally applicable methodology to assess the performance of realized measures when the variable of interest is latent,overcoming the problem posed by the lack of a true value for the integrated variance. We use E-mini index futures contracts to carry out formal testing of several realized measures in the presence of noise. Moreover, a thorough simulation analysis is employed to evaluate the estimators' sensitivity to different price and noise processes, and sampling frequencies. These results provide clear and valuable insights on the strengths and weaknesses of the analysed estimators.

C1461: **Testing the correct specification of a spatial panel model for stock returns**
Presenter: **Tim Kutzker**, University of Cologne, Germany
Co-authors: Dominik Wied
A test for spatial autoregression models is provided. Spatial autoregression models usually incorporate global dependencies, dependencies inside industrial branches and local dependencies. The power of the test is also shown. An empirical analysis of the Euro Stoxx 50 returns concludes the findings.

C1574: **Multivariate specification tests based on a dynamic Rosenblatt transform**
Presenter: **Igor Kheifets**, ITAM, Mexico
Parametric model adequacy tests for multivariate dynamic models are considered. We show that commonly used Kolmogorov-type tests do not take into account cross-sectional nor time dependence structure and propose a test based on multi-parameter empirical processes that overcomes these problems. We propose a simulation experiment to illustrate the properties of the tests. The tests are applied to a LSTAR-type model of joint movements of the UK output growth and interest rate spreads. Asymptotic properties of the test statistics under the null of correct specification and under the local alternative, and justification of a parametric bootstrap to obtain critical values are provided.

CG082 Room MAL 151 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I Chair: Fulvio Corsi

C0280: **The information content of short-term options**
Presenter: **Lazaros Symeonidis**, University of East Anglia, United Kingdom
Co-authors: Ioannis Oikonomou, Andrei Stancu, Chardin Wese
It is documented that the implied variance of daily and weekly maturities strongly predict next month's realized variance. We introduce the HAR-IV model that jointly uses the daily, weekly and monthly implied variance to predict realized variance. The HAR-IV model outperforms the HAR-RV model both in- and out-of-sample. An investor would pay up to 3.887% per year to switch from the timing strategy based on the HAR-RV model to the strategy based on the HAR-IV model. Our results are robust to heteroscedastic measurement errors.

C1514: **Realized volatility modelling with measurement errors and nonlinear effects**
Presenter: **Fulvio Corsi**, Ca Foscari University Venice and City University London, Italy
Co-authors: Giuseppe Buccheri
Despite its effectiveness, the approximate long-memory HAR model neglects measurement errors and exhibits several evidences of misspecification due to the inherent nonlinearity of the realized volatility dynamics. We propose new extensions of the HAR model apt to address these effects

separately with the aim to disentangle them and quantify their contribution in improving volatility forecasts. First, we combine the asymptotic theory of the realized volatility estimator with Kalman filter to account for measurement errors. Secondly, nonlinear effects are captured by introducing time variations in the HAR parameters driven by the score of the predictive likelihood. The two approaches are then combined to simultaneously account for both measurement errors and nonlinearities. The proposed models are simply estimated through standard maximum likelihood methods and are shown, both on simulations and on real data, to provide better out-of-sample volatility forecasts compared to existing approaches.

C1207: **Global bond market interaction using regime-switching dynamic term structure model**
Presenter: **Takeshi Kobayashi**, Nagoya University of Commerce & Business, Japan
The globalization process has significant impact on the world bond market. Global term structure dynamics are investigated by using a hierarchical factor model. We examine advanced countries and developing countries during 2003-2017 which focus on Asian sovereign bonds market which have increased steadily. We use a two-step state space model to extract the global factor. The results indicates yield curves of some countries possess their own dynamic while the other countries yield curves are strongly influenced by global level and slope factors. Regime switching extension of the global yield curve model are also developed. In this model the volatilities of global factor shift switch between two regimes. A previous method is implemented to construct likelihood function and unknown parameters. We show how the linkage between the global factor and country specific factors differs across two bond market volatility regimes.

C1532: **On speculative bubbles in the Chinese stock market: Evidence from cross-listed stocks**
Presenter: **Efthymios Pavlidis**, Lancaster University Management School, United Kingdom
Co-authors: Kostas Vasilopoulos
During the last few decades, the Chinese stock market has experienced great turbulence, with periods of rapid price increases followed by severe market crashes. A widely held view is that this volatile behavior reflects speculative bubbles which drive asset prices away from their intrinsic values and then burst. Although popular, this hypothesis is difficult to examine. The difficulty lies in the fact that the intrinsic value of financial assets is not directly observable. As a consequence, most econometric tests for bubbles actually examine a composite hypothesis: no bubbles and a correctly specified model for market fundamentals. Because rejection of the null may be solely due to model misspecification, such tests are inconclusive. We propose a novel approach that circumvents this obstacle by utilizing information incorporated in the share prices of Chinese cross-listed companies. The basic idea is that an A-share that is listed on the mainland Chinese stock market and an H-share that is listed on the Hong Kong Stock Exchange by the same company have the same underlying fundamentals and, therefore, the presence of explosive dynamics in their price difference provides conclusive evidence of speculative bubbles. By applying a recursive right-tailed unit root tests as well as a wild-bootstrap version of it, we find that a substantial number of cross-listed companies display speculative bubbles.

C1552: **Extracting latent states from high frequency option prices**
Presenter: **Genevieve Gauthier**, HEC Montreal, Canada
Co-authors: Jean-Francois Begin, Diego Amaya
The aim is to propose the realized option variance as a new observable variable to integrate high frequency option prices in the inference of option pricing models. Using simulation and empirical studies, the incremental information offered by this realized measure is documented. Our empirical results show that the information contained in the realized option variance improves the inference of model variables such as the instantaneous variance and variance jumps of the S&P 500 index. Parameter estimates indicate that the risk premium breakdown between jump and diffusive risks is affected by the omission of this information.

CP001 Room Macmillan Hall and Crush Hall POSTER SESSION Chair: Marios Fyrillas

C0946: **Bayesian approach to the multivariate Fay-Herriot model**
Presenter: **Tsubasa Ito**, University of Tokyo, Japan
Small area estimation deals with inference problems for small areas with small sample sizes. In this case, direct design based estimators for small domains can be improved by incorporating relevant supplementary information available from administrative records through linking models. Linear mixed models are often used and these models use random area effects for the between area variation of the data, which is not explained by these supplementary information. In small area estimation, the Fay Herriot model is widely used. This model is an area level linear mixed model with random area effects. We consider multivariate Fay Herriot models for small area estimation. Statisticians or administrators are often required to estimate multiplicate indicators, such as income, poverty rate, unemployment rate, health expenditure and so on. In this case, the performance of the estimators are improved by taking into account for the correlation of these variables rather than estimating each indicator respectively. We set the covariance matrix of random effects is general form, which has not been studied in existing papers, derive the empirical best predictor of the vector of area means and give an approximation of the matrix of mean squared cross prediction error. Moreover, we show the risk of prediction can be improved by using a spike and slab prior as a prior density for random effects, which is a phenomenon peculiar to the multivariate case.

C1580: **A statistical analysis of uncertainty for conventional and ethic investments**
Presenter: **Nabila Jawadi**, IPAG Business School, France
Co-authors: Abdoulkarim Idi cheffou, Fredj Jawadi
The aim is to estimate and compare uncertainty for two classes of ethic shares (Islamic stock Market and Social and Responsible Investments) with reference to the US conventional stock market. To this end, we use the class of GARCH modeling to measure uncertainty for three major stock indexes (Dow Jones Industrial Index, Dow Jones Islamic Index, Dow Jones Sustainable Index) over the period 1996-2017. Our analysis points to significant difference of uncertainty level across conventional and ethic investments that varies with market state (calm period, crisis period). These findings might have implications for investors to adjust and optimize their portfolio composition.

C1778: **Effects of some indicators on existence of anomalies in frontier stock markets**
Presenter: **Anna Czapkiewicz**, AGH University of Science and Technology, Poland
Co-authors: Adam Zaremba
The study discusses the performance of 120 anomalies from the finance literature in the frontier markets. Taking into account properties of financial time series, we discuss the Markov switching AR(1) model with two states to verify the anomaly existence. Furthermore, we consider the role of some indicators for the anomaly occurrence. We verify the effects of the indicators of investor sentiment and limits on arbitrage. For this purpose, we adapt the Markov switching AR(1) model with time-varying matrix transition probability (TVTMP). We demonstrate convincing evidence for the predictability in anomaly performance.

C1664: **Asset pricing in the quantile-frequency domain**
Presenter: **Matej Nevrla**, Charles University, Czech Republic
Co-authors: Jozef Barunik
Despite many attempts to understand cross-section of asset returns, there is no consensus on the functional form of the pricing kernel, cornerstone

of the asset pricing theory. Many simplifying assumptions, such as quadratic preferences in CAPM model, led to results not supported by the real-world data. In the recent literature, two fruitful approaches emerged trying to explain risk premium of the assets. One with emphasis on asymmetric features of the asset returns based on the notion that agents put more weight on downside risk then upside potential when pricing an asset. The other approach stresses that risk aggregated over all investment horizons is not an adequate measure and emphasises the importance of frequency-specific risk and its implications for agents caring differently over various economic cycles. We aim to unify these two approaches and show that downside risk possesses complex structure and differs over various cycles. We do this by defining quantile cross-spectral betas that capture exposure to the downside risk over different investment horizons. Downside risk in our approach is characterized by the quantile dependency, which is further decomposed into frequency domain. We apply it to the wide range of asset classes and asses how the model jointly prices them.

C1681: **Dynamic quantile models, rational inattention, and asset prices**
Presenter: **Lukas Vacha**, Univerzita Karlova, Fakulta socialnich ved, Czech Republic
Co-authors: Jozef Barunik
The aim is to study asset pricing under uncertainty with agents having quantile preferences, and limited information processing capacity. Abandoning the classical asset pricing that relies on expected utility, we introduce a dynamic quantile model for asset pricing, in which the agent maximizes stream of future quantile utilities instead. In addition, agent cannot acquire all information about future states of her portfolio freely. In contrast to the rational expectation models, the agent has limited amount of attention since the information she obtains are costly. In our model, the agent maximizes stream of her future quantile utilities according to her quantile utility preferences subject to information costs constraints. Our results show that there is a significant benefit when standard expected utility is expanded into quantile preference utilities.

C1615: **Frontier markets efficiency: Mutual information and DFA analyses**
Presenter: **Andreia Dionisio**, University of Evora, Portugal
Co-authors: Wahbeeah Mohti, Isabel Vieira
The weak form efficiency in frontier markets is investigated. We analyze stock market indexes of 23 countries and use mutual information and detrended fluctuation analysis (DFA) to examine serial dependence and non linear dependence in the series. The results from the former approach indicate the existence of non-linearity in the return series. Results obtained with DFA suggest that the return of several frontier markets including Sri Lanka, Bulgaria, Serbia, Estonia, Mauritius and Kazakhstan have more pronounced long-term dependence and exhibit persistent behavior. The series of some frontier markets including Argentina, Lebanon, Bahrain, Tunisia, Bangladesh and Vietnam, display long-term dependence with anti-persistent behavior. The obtained results indicate that the assessed frontier markets are weak form inefficient.

C0330: **Modelling persistence in the conditional mean of inflation using an ARFIMA process with GARCH and GJR-GARCH innovations**

Presenter: **Alexander Boateng**, University of Limpopo, South Africa
Co-authors: Luis Alberiko Gil-Alana
The aim is to contribute to the debate on inflation persistence by extending an ARFIMA process with GARCH and GJR-GARCH models to describe the time-dependent heteroscedasticity and persistence in the conditional mean of Consumer Price Index (CPI) inflation series of Ghana and South Africa under three distributional assumptions (i.e., Normal, Student-t and Generalised Error Distributions). ARFIMA(3,0.26,1)-GJR-GARCH(1,1) under Generalised Error Distribution and ARFIMA(3,50,2)-GJR-GARCH(1,1) under Student-t Distribution respectively, provided the best fit for modelling the time-dependent heteroscedasticity and persistence in the conditional mean of CPI inflation rate of Ghana and South Africa. Results from the study provided evidence of persistence, mean reverting though, and asymmetric effect of economic shocks on the conditional mean of CPI inflation rate of the two countries. These results would, therefore, be useful to both countries in making good portfolio decisions, accessing the efficacy of a monetary policy or programme meant to control inflation persistence and also serving as a tool for detecting volatility and its impact, in the Ghanaian and South African inflation rate and their economies at large.

C1857: **Robust on-line portfolio selection via adaptive conditional volatility estimation**
Presenter: **Minyoung Kim**, Seoul National University of Science and Technology, Korea, South
In the on-line portfolio selection problem, the principle of the moving average reversion (MAR) has received significant attention recently. However, most MAR-based approaches focused on the point estimates of the MAR, unable to deal with its volatility, leading to non-robust strategies. We propose a quite simple but reasonable volatility model: a conditional volatility model where the volatility of the return is assumed to be a function of the MAR estimate. It is motivated from the observation that the volatility tends to be high when the MAR estimate is positively large, and vice versa, which is indeed verified in several market historic data. We specifically model this phenomenon as a simple step function comprised of two sets of parameters, the MAR threshold and the levels of the volatility. These parameters are adaptively estimated from the latest data via Bayesian inference where we place conjugate priors (e.g., scaled-inverse-chi2) on them. With this volatility model, we optimize the portfolio using the worst-case Value-At-Risk method, which can be formulated as an SDP optimization that admits many efficient solvers. Additionally, the method becomes more robust since it makes no assumption on specific distributions (e.g., Gaussians) beyond the second-order moment constraints. We empirically demonstrate the benefits of the proposed approach on various real stock market data.

C1856: **On the cross-influence of cryptocurrencies and traditional asset classes**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic
Large stream of literature studies interconnectedness among various assets that are relevant in current global markets. Transmission of shocks between cryptocurrencies and traditional asset classes is, however, not understood at all, but should not be ignored due to increasing influence of cryptocurrencies in recent years. We study how shocks between the most liquid representatives of the traditional asset classes including commodities, foreign exchange, stocks, financials, and cryptocurrencies are being transmitted. Generally, we document very low level of connectedness between the main cryptocurrency and other studied assets. The only exception is gold which receives substantial amount of shocks from cryptocurrency market. Our findings are important since we show that cryptocurrencies play role in global markets, and the results could also be useful in portfolio diversification schemes. Moreover, we find significant positive asymmetry in spillovers between the studied assets, which is in contradiction to previous studies conducted on assets from a single asset class.

C1864: **Portfolio diversification in the spectral domain**
Presenter: **Martin Hronec**, UTIA AV CR, Czech Republic
Co-authors: Jozef Barunik
When investors' risk preferences differ across time horizons, the diversified portfolio needs to be immune not only against shocks aggregated across time horizons but also at specific time horizons. We apply spectral analysis into diversification-based portfolio selection models. By replacing the covariance matrix estimates with the cross-spectrum based ones, we restrict the optimization problems in these models to the desired frequency band, allowing an investor to target specific time horizons. Further, we generalize for an investor facing risk constraints at different frequencies by including the shape of the cross-spectrum into the optimization problem. We provide several numerical and empirical examples that show investors may benefit by considering not only diversification across aggregate risk sources but also across different frequencies.

Sunday 17.12.2017 14:25 - 16:05 Parallel Session I – CFE-CMStatistics

EO208 Room CLO B01 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING II **Chair: Matteo Ruggiero**

E1346: **Non-exchangeable random partition model for microclustering**
Presenter: **Francois Caron**, University of Oxford, United Kingdom
Co-authors: Giuseppe Di Benedetto, Yee Whye Teh
Clustering aims at finding a partition of the data. In a Bayesian framework, this task is addressed by specifying a prior distribution on the partition of the data. Popular models, such as the Chinese Restaurant Process and its two-parameters generalization, rely on some exchangeability assumption; while this assumption may be reasonable for some applications, it has strong implications on the asymptotic properties of the cluster sizes. In fact, exchangeable random partitions imply the linear growth of the cluster sizes, which is not suitable for several applications. We will present a flexible non-exchangeable random partition model, based on completely random measures, which is able to generate partitions whose growth of the clusters sizes is almost surely sublinear. Along with this result, we provide the asymptotic behaviour of the number of clusters and of the proportion of clusters of a given size. Sequential Monte Carlo algorithms are derived for inference and we provide an illustration of the fit of the model on a movie review dataset.

E0927: **Exploiting conjugacy to build time dependent completely random measures**
Presenter: **Ilaria Bianchini**, Politecnico di Milano, Italy
Co-authors: Raffaele Argiento, Jim Griffin
A flexible approach to build stationary time-dependent processes exploits the concept of conjugacy in a Bayesian framework: in this case, the transition law of the process is defined as the predictive distribution of an underlying Bayesian model. Then, if the model is conjugate, the transition kernel can be analytically derived, making the approach particularly appealing. We aim at achieving such a convenient mathematical tractability in the context of completely random measures (CRMs), i.e. when the variables exhibiting a time dependence are CRMs. In order to take advantage of the conjugacy, we consider the wide family of exponential completely random measures. This leads to a simple description of the process which has a autoregressive structure of order 1. The proposed process can be straightforwardly employed to extend CRM-based Bayesian nonparametric models such as feature allocation models to time-dependent data. These processes can be applied to problems from modern real life applications in very different fields, from computer science to biology. In particular, we develop a dependent latent feature model for the identification of features in images and a dynamic Poisson factor analysis for topic modelling, which are fitted to synthetic and real data.

E1114: **Correlated random measures**
Presenter: **Rajesh Ranganath**, Princeton University, United States
Co-authors: David Blei
Many hierarchical Bayesian nonparametric models are built from completely random measures, in which atom weights are independent. This leads to implicit independence assumptions in the corresponding hierarchical model, assumptions that are often misplaced in real-world settings. We address this limitation. We develop correlated random measures, a class of random measures where the measures on two disjoint sets can exhibit both positive and negative dependence. Correlated random measures model correlation within the measure by using a Gaussian process in concert with a Poisson process. With this construction, for example, we can develop a latent feature model for which we can infer both the properties of the latent features and their correlation. We develop several examples of correlated random measures and touch on correlated Poisson Kingman constructions. We show improved predictive performance on large collections of text and large collections of medical diagnostic codes.

E0828: **Investigating predictive probabilities of Gibbs-type priors**
Presenter: **Julyan Arbel**, Inria, France
Co-authors: Stefano Favaro
Gibbs-type priors are arguably the most 'natural' generalization of the Dirichlet prior. Among them the two parameter Poisson-Dirichlet prior certainly stands out for the simplicity and intuitiveness of its predictive probabilities. Given an observable sample of size n, we show that the predictive probabilities of any Gibbs-type prior admit a large n approximation, with an error term vanishing as $o(1/n)$, which maintains the same mathematical tractability and interpretability as the predictive probabilities of the two parameter Poisson-Dirichlet prior. We discuss the use of our approximate predictive probabilities in connection with some recent work on Bayesian nonparametric inference for discovery probabilities.

EO740 Room MAL B18 GRAPHICAL MARKOV MODELS III **Chair: Elena Stanghellini**

E1288: **A generic algorithm for estimation in undirected graphical models**
Presenter: **Steffen Lauritzen**, University of Copenhagen, Denmark
A generic and globally convergent algorithm is presented for estimation in undirected graphical models with only pairwise interaction terms which can include penalty functions of lasso type, total positivity restrictions, and specific restrictions on the graph type. The algorithm is essentially a variant of Iterative Proportional Scaling and contains the latter as a special instance.

E1063: **Markov and other properties of cyclic structural causal models**
Presenter: **Joris Mooij**, University of Amsterdam, Netherlands
Structural Causal Models (SCMs), also known as (Non-Parametric) Structural Equation Models (NP-SEMs), are widely used for causal modelling purposes. One of their advantages over other representations such as causal Bayesian networks is that SCMs allow for cycles (causal feedback loops). The presence of cycles adds many complexities that are absent in the acyclic setting, especially for nonlinear models. We will discuss recent advances in the theory of cyclic Structural Causal Models. We will discuss how they can be marginalized to describe a subsystem of interest, explain how SCMs can model the equilibrium states of ordinary differential equation models and how these equilibrium states change under perturbations, and we will present recent results on the Markov properties of cyclic SCMs that provides the corner stone for novel cyclic causal discovery algorithms.

E1195: **Graphical models based on trees**
Presenter: **Anna Gottard**, University of Firenze, Italy
Graphical models have been utilised in a wide range of problems to characterise the conditional independence structure among random variables. Particularly interesting are applications in genomics and omics science. A better understanding of the association among gene/protein/metabolite molecular signatures potentially offers new insights for complex diseases. With continuous random variable, most of this research traditionally focuses on Gaussian graphical models, assuming linear relationships. However, the assumptions of multivariate Gaussianity and linearity in the dependence structure are often evidently erroneous. Recent literature explores graphical models with non-linear relations. We investigate pairwise graphical models on a set of random variables, with distributions in which dependence occurs through the expected value. We study the utility of tree based models to detect interactions and non-linearities in these distributions and compare different algorithm for searching a tree or a sum of

trees. A particular case of quasi-linear systems is analysed, embedding linear and nonlinear effects.

EO210 Room MAL B20 MULTIVARIATE EXTREME VALUE Chair: Armelle Guillou

E0247: **Tail dependence measure for examining financial extreme co-movements**
Presenter: **Vali Asimit**, City University London, United Kingdom
Co-authors: Russell Gerrard, Yanxi Hou, Liang Peng
Modeling and forecasting extreme co-movements in financial market is important for conducting stress test in risk management. Asymptotic independence and asymptotic dependence behave drastically different in modeling such co-movements. For example, the impact of extreme events is usually overestimated whenever asymptotic dependence is wrongly assumed. On the other hand, the impact is seriously underestimated whenever the data is misspecified as asymptotic independent. Therefore, distinguishing between asymptotic independence/dependence scenarios is very informative for any decision-making and especially in risk management. We investigate the properties of the limiting conditional Kendall's tau which can be used to detect the presence of asymptotic independence/dependence. We also propose non-parametric estimation for this new measure and derive its asymptotic limit. A simulation study shows good performances of the new measure and its combination with a coefficient of tail dependence previously proposed. Finally, applications to financial and insurance data are provided.

E0731: **Inference methods for dependent and asymptotically independent extremes**
Presenter: **Stefano Rizzelli**, Bocconi University, Italy
Co-authors: Simone Padoan, Armelle Guillou
Inferring the dependence among extreme observations is a crucial aim of multivariate extreme-value analysis. Classical extreme-value theory provides asymptotic models for the tail probability of multivariate distributions that are in the domain of attraction of the so-called multivariate extreme value distribution. The latter is the limiting distribution of the normalized component-wise maxima. When the attractor is a product of univariate extreme value margins, a multivariate distribution is said asymptotically independent and the corresponding model for the tail probability is degenerate. In the last few decades, several characterizations of this phenomenon have emerged, and a new theory on tail probabilities has been established. Such theoretical developments are particularly useful for analyzing data that exhibit a positive association which, however, is mitigated at more and more extreme levels. Within this theoretical framework, a new dependence function is introduced which parallels the classical Pickands function but suits asymptotically independent extremes. A semi-parametric estimator of it is proposed and its asymptotic properties are established. Moreover, a statistical test for asymptotic independence is developed, which is based on the classical Pickands dependence function and suits data dimensions larger than two. The performances of the proposed inferential methods are illustrated by simulation studies.

E0728: **The Kullback-Leibler divergence in testing multivariate extreme value models**
Presenter: **Chen Zhou**, Erasmus University Rotterdam, Netherlands
Co-authors: Sebastian Engelke, Philippe Naveau, Chen Zhou
Many effects of climate change are reflected in the frequency and severity of the extreme events in the tail of the distributions. Detecting such changes requires a statistical methodology that can test the distributional changes in the large observations in the sample. We propose a simple, non-parametric test that decides whether two multivariate distributions exhibit the same tail behavior. The test is based on the Kullback-Leibler divergence, between exceedances over a high threshold of the two multivariate random vectors. We show that such a divergence measure is closely related to the divergence between Bernoulli random variables. We study the properties of the test and further explore its effectiveness for finite sample sizes. As an application we apply the method to precipitation data where we test whether the marginal tails and/or the extreme value dependence structure have changed over time.

E0747: **Limits to human life span through extreme value theory**
Presenter: **John Einmahl**, Tilburg University, Netherlands
Co-authors: Jesson Einmahl, Laurens de Haan
There is no scientific consensus on the fundamental question whether the probability distribution of the human life span has a finite endpoint or not and, if so, whether this upper limit changes over time. Crucially, the limit is not defined as the highest observed age at death but the highest age that possibly could be reached. Our approach to the problem is to concentrate on precisely observed mortality data. The study is based on a unique dataset of the ages at death in days of all Dutch residents, born in the Netherlands, who died in the years 1986-2015 at a minimum age of 92 years. Unlike most other studies we use extreme value theory and base our analysis on the configuration of thousands of mortality data of old people, not just the few oldest old. The existence of a finite upper limit to the life span will follow from the fact that the 30 annual extreme value indices (not only their estimates) take on only negative values. We find compelling statistical evidence that there is indeed an upper limit to the life span of men and to that of women for all the 30 years we consider and, moreover, that there are no indications of trends in these upper limits over the last 30 years, despite the fact that the number of people reaching high age (say 95 years) is increasing rapidly, almost tripling, in these 30 years. Using extreme value theory, we also present estimates for the endpoints and for the force of mortality at high age.

EO356 Room MAL B30 TIME SERIES AND EXTREMES Chair: Thomas Mikosch

E1241: **On records of stationary heavy tailed sequences**
Presenter: **Bojan Basrak**, University of Zagreb, Croatia
Records and records times of stationary regularly varying sequences are studied. Using point processes theory adapted to some non standard spaces, we are able to describe asymptotic distribution for the extremes and records in such a sequence, as long as the dependence between the observations is sufficiently weak. In particular, we obtain a remarkably simple structure for the limiting distribution of the record times. We also discuss possible extensions of these results to spatio temporal data.

E0858: **A comparison of high-dimensional sample covariance and correlation matrices of a heavy-tailed time series**
Presenter: **Johannes Heiny**, Aarhus University, Denmark
Co-authors: Thomas Mikosch
In Principal Component Analysis one studies the sample covariance or sample correlation matrix, both of which often lead to the same result. We first analyze the joint distributional convergence of the largest eigenvalues of the sample covariance matrix of a high-dimensional heavy-tailed time series. Assuming a regular variation and infinite fourth moment, we employ a large deviations approach to show that the extreme eigenvalues are essentially determined by the extreme order statistics from an array of iid random variables. The asymptotic behavior of the extreme eigenvalues is then derived routinely from classical extreme value theory. The resulting approximations are strikingly simple considering the high dimension of the problem at hand. Then we compare the behavior of the eigenvalues of the sample covariance and sample correlation matrices and argue that the latter seems more robust, in particular in the case of infinite fourth moment. We show that the largest and smallest eigenvalues of a sample correlation matrix stemming from n independent observations of a p-dimensional time series with iid components converge almost surely to constants, as $p,n \to \infty$, if the truncated variance of the entry distribution is 'almost slowly varying', a condition we describe via moment

properties of self-normalized sums. Moreover, the empirical spectral distributions of these sample correlation matrices converge weakly to the Marchenko-Pastur law.

E0659: **Limit theorems for empirical cluster functionals with applications to statistical inference**
Presenter: **Rafal Kulik**, University of Ottawa, Canada
Limit theorems for empirical cluster functionals are discussed. Conditions for weak convergence are provided in terms of tail and spectral tail processes and can be verified for a large class of multivariate time series, including geometrically ergodic Markov chains. Applications include asymptotic normality of blocks and runs estimators for the extremal index and other cluster indices. Results for multiplier bootstrap processes are also provided.

E0633: **Heavy tails for an alternative stochastic perpetuity model**
Presenter: **Olivier Wintenberger**, Sorbonne University, France
Co-authors: Thomas Mikosch, Mohsen Rezapour
A stochastic model of perpetuity-type is considered. In contrast to the classical affine perpetuity model, all discount factors in the model are mutually independent. We prove that the tails of the distribution of this model are regularly varying both in the univariate and multivariate cases. Due to the additional randomness in the model the tails are not pure power laws as in the classical setting, but involve a logarithmic term.

EO670 Room MAL B33 RECENT DEVELOPMENTS FOR COMPLEX DATA Chair: Juan Romo

E0328: **Spectra-based clustering methods for visualizing spatio-temporal patterns of winds and waves in the Red Sea**
Presenter: **Carolina Euan**, King Abdullah University of Science and Technology, Saudi Arabia
In oceanic research, it is challenging to understand the patterns of winds and waves due to the complicated spatio-temporal dynamics. We propose new spectra-based methods for clustering hourly data of wind speed and wave height observed in the entire Red Sea. By clustering time series observed from different locations together, we identify spatial regions that share similar wind and wave directional spectra. We show that it is necessary to consider directional spectra for winds and waves, and that the clustering results may be very different, ignoring the direction. Finally, we develop an application to visualize the resulting time-evolving clusters in the Red Sea with a chosen clustering algorithm.

E0947: **A bootstrap approach to the inference on dependence in a multivariate functional setting**
Presenter: **Francesca Ieva**, Politecnico di Milano, Italy
Co-authors: Juan Romo, Francesco Palma
An inferential bootstrap-based procedure is presented for the of Spearman index, i.e., an index which aims to quantify the level of dependence between two families of functional data. We provide point and interval estimators of the index in order to check, through suitable tests, if two families of functional data can be considered as being independent. We introduce the new notion of Spearman Matrix (SM), which enables us to describe the pattern of dependence among the components of a multivariate functional dataset. A simulation study aimed at testing the performance of the Spearman index and matrix in correctly detecting the dependence is also provided. Finally, SM is used to analyze two different populations of multivariate curves (specifically, Electrocardiographic signals of healthy and unhealthy people), in order to check if the patterns of dependence between the components are different in the two cases. This is done by providing the results of suitable hypothesis tests verifying the equality between the Spearman matrices of the two populations.

E0598: **The ELFE survey and its complex survey design**
Presenter: **Anne Ruiz-Gazen**, Toulouse School of Economics, France
The 2011 French Longitudinal Survey on Childhood ELFE comprises more than 18,000 children selected on the basis of their place and date of birth. A sample of 320 maternity units and a sample of 25 days divided in four time periods have been selected. The sample is made of babies born at the sampled locations and on the sampled days. The objective of this large survey is to analyze the physical and psychological health of children together with their living and environmental conditions. In order to derive reliable confidence intervals for finite population parameters, the ELFE sampling design has to be taken into account. This design consists in drawing samples from a two-dimension population, independently in each dimension, and is called the Cross-Classified Sampling design. A general theory of estimation is derived for this design. Generally, the cross-classified design results in a loss of efficiency as compared to the usual two-stage design. Even in the case of simple random sampling designs without replacement in each dimension, the usual variance estimators may be negative. Non-negative simplified variance estimators are introduced and compared through a simulation study. Finally, an application to the ELFE data is detailed.

E0778: **Functional variable selection based on RKHS**
Presenter: **Jose Luis Torrecilla**, Universidad Carlos III de Madrid, Spain
Co-authors: Jose Berrendero, Antonio Cuevas
Variable selection techniques have become a popular tool for dimension reduction with an easy interpretation. However, we are still far from getting a standard in the functional classification framework. We propose a new functional-motivated variable selection methodology (RK-VS). This method appears as a direct consequence of looking at the functional classification problem from an RKHS (Reproducing Kernel Hilbert Space) point of view. In this context, under a general Gaussian model and a sparsity assumption, the optimal rules turn out to depend on a finite number of variables. These variables can be selected by maximizing the Mahalanobis distance between the finite-dimensional projections of the class means. Our RK-VS method is an iterative approximation to this. This is an easy-to-interpret and fast methodology which allows for easily adding extra information about the model. The empirical performance of RK-VS is extremely good when the considered problems fit the assumed model but it turns out to be also quite robust against partial departures from the hypotheses, typically leading to very good results in general problems.

EO555 Room MAL B34 HIGH-DIMENSIONAL FUNCTIONAL DATA ANALYSIS Chair: Matthew Reimherr

E1012: **Boosting functional response models for location, scale and shape with an application to bacterial competition**
Presenter: **Almond Stoecker**, LMU Munich, Germany
Co-authors: Sarah Brockhaus, Sophia Schaffer, Benedikt von Bronk, Madeleine Opitz, Sonja Greven
Generalized Additive Models for Location, Scale, and Shape (GAMLSS) are extended to regression with functional responses. GAMLSS are a flexible model class allowing for modeling multiple distributional parameters at once. The model is fitted via gradient boosting, which provides inherent model selection. We apply the functional GAMLSS to analyze bacterial interaction in Escherichia coli and show how the consideration of variance structure fruitfully extends usual growth models.

E0706: **Non-parametric multi-aspect local null hypothesis testing for functional data: Analysis of articulatory phonetics data**
Presenter: **Simone Vantini**, Politecnico di Milano, Italy
Co-authors: Alessia Pini, Lorenzo Spreafico, Alessandro Vietti
The focus is on the statistical comparison of ultrasound tongue profiles pertaining to different allophones pronounced by the same speaker which

can be modelled as functions varying on a spatio-temporal domain. Stimulated by this application we will introduce a general framework for multi-aspect local non-parametric null-hypothesis testing for functional data. In detail: "multi-aspect" pertains to the fact the procedure allows the simultaneous investigation of many different data aspects like means and variances of tongue vertical position, slope, concavity, velocity, and acceleration; "local" pertains instead to the fact the procedure can impute the rejection to aspect-specific regions of the domain; finally, "non-parametric" refers to the fact that the specific implementation of the procedure is permutation-based and thus finite-sample exact and consistent independently on data Gaussianity. For ease of clarity, the focus will be on functional two-population tests and ANOVA. Nevertheless, the approach is flexible enough to be adapted to more complex testing problems like functional linear regression.

E1245: **Some theoretic results on functional data analysis with penalized splines**
Presenter: **Luo Xiao**, North Carolina State University, United States

Penalized spline methods are popular for analyzing functional data. However, the theoretical foundation of penalized splines in functional data analysis is largely unknown. We attempt to fill the theoretical gap and introduce some new theoretical results in the estimation of the mean and covariance functions via penalized splines. In particular, we show that penalized splines are rate optimal and have two-type asymptotics similar to previous results for univariate penalized splines in the context of nonparametric regression.

E1595: **Probabilistic K-mean with local alignment for functional motif discovery**
Presenter: **Marzia Cremona**, The Pennsylvania State University, United States
Co-authors: Francesca Chiaromonte

The aim is to address the problem of discovering functional motifs, i.e. typical "shapes" that may recur several times in a set of (multidimensional) curves, capturing important local characteristics of these curves. We formulate probabilistic K-mean with local alignment, a novel algorithm that leverages ideas from Functional Data Analysis (joint clustering and alignment of curves), Bioinformatics (local alignment through the extension of high similarity "seeds") and fuzzy clustering (curves belonging to more than one cluster, if they contain more than one typical "shape"). Our algorithm identifies shared curve portions, which represent candidate functional motifs in a set of curves under consideration. It can employ various dissimilarity measures in order to capture different shape characteristics. After demonstrating the performance of the algorithm on simulated data, we apply it to discover functional motifs in "Omics" signals related to mutagenesis and genome dynamics, exploring high-resolution profiles of different mutation rates in regions of the human genome where these rates are globally elevated.

EO672 Room MAL B35 UNDERSTANDING BRAIN FUNCTIONAL CONNECTIVITY Chair: Hernando Ombao

E0391: **Gradient synchronization as a measure for brain functional connectivity**
Presenter: **Jane-Ling Wang**, University of California Davis, United States
Co-authors: Hans-Georg Mueller, Yang Zhou, Owen Carmichael

Quantifying the association between components of multivariate signals is a key to assess brain functional connectivity. Recent research suggests temporally changing patterns of functional connectivity. We propose new similarity measures for pairs of temporal functions that reflect the dynamic features of their functional similarity. We introduce a gradient synchronization measure that is based on the aggregated concordance and discordance of the gradients between paired smooth random functions, which can also be used to quantify the relative frequency of changes from concordance to discordance or vice versa. Asymptotic normality of the proposed estimates is obtained under regularity conditions and the new methods are illustrated via simulations and an application to resting state fMRI signals from Alzheimer's patients and healthy subjects.

E1409: **Estimating dynamic connectivity states in fMRI**
Presenter: **Chee Ming Ting**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Hernando Ombao

Sliding-window analysis or time-varying coefficient models which are unable to capture both smooth and abrupt changes simultaneously in dynamic brain connectivity. Emerging evidence also suggests state-related changes in brain connectivity where dependence structure alternates between a finite number of latent states or regimes. Another challenge is to infer full-brain networks with large number of nodes. A novel approach is proposed based on a regime-switching factor model. The dynamic connectivity states in high-dimensional brain signals is characterized via a few common latent factors. By assuming the factor dynamics to follow Markov-switching vector autoregressive (VAR) process, regime-switching in high-dimensional directed dependence between the observations is allowed. This model enables efficient, data-adaptive estimation of change-points of effective brain connectivity regimes and the massive dependencies associated with each regime. We introduce a two-step estimation procedure: (1) Extract the factors by PCA and (2) Estimate the switching models in a state-space form by maximum likelihood method using the Kalman filter the EM algorithm. We also used the wavelet-based functional mixed models to analyze dynamic connectivity states that are varied across multiple subjects or multiple signal replicates. The method is applied in multi-subject resting-state fMRI data where mental activities are unconstrained, and in multi-trial EEGs evoked by repetitive auditory stimuli.

E1736: **Fréchet estimation of dynamic covariance matrices, with application to regional myelination in the developing brain**
Presenter: **Alexander Petersen**, University of California Santa Barbara, United States
Co-authors: Hans-Georg Mueller, Sean Deoni

Assessing brain development for small infants is important for determining how the human brain grows during the early period of life when the growth rate is at its peak. MRI techniques enable the quantification of brain development, with a key quantity being the level of myelination, where myelin acts as an insulator around nerve fibers and its deployment makes nerve pulse propagation more efficient. The co-variation of myelin deployment across different brain regions provides insights into the co-development of brain regions and can be assessed as correlation matrix that varies with age. Typically, available data for each child are very sparse, due to the cost and logistic difficulties of arranging MRI brain scans for infants. We showcase here a method where data per subject are limited to measurements taken at only one random age, so that one has cross-sectional data available, while aiming at the time-varying dynamics. The challenge is that at each observation time one observes only a p-vector of measurements but not a covariance or correlation matrix. For such very sparse data, we develop a Fréchet estimation method that generates a matrix function where, at each time, the matrix is a non-negative definite covariance matrix, for which we demonstrate consistency properties. We discuss how this approach can be applied to myelin data in the developing brain and what insights can be gained.

E1644: **Time varying brain functional connectivity: Change-point estimation and testing**
Presenter: **DuBois Bowman**, Columbia University, United States
Co-authors: Jaehee Kim, Bin Cheng, Daniel Drake

Functional connectivity (FC) methods are widely used to explore alterations in interregional associations that are linked to disease processes such as major depressive disorder (MDD) and Parkinson's disease (PD). There is emerging evidence toward time varying patterns of FC, although the vast majority of studies consider only static scalar-valued summaries of FC between pairs of brain regions. Many current approaches to dynamic FC are highly sensitive to a basic assumption regarding the choice of window length. We propose a CUSUM (or cumulative sum) analysis framework to assess time varying FC, with estimation and testing based on temporal profiles of scalar-valued functions of the sample covariance matrix. We examine the performance of our method using simulation studies and apply our method to determine change points in functional magnetic

resonance imaging (fMRI) data from patients and healthy controls.

EO682 Room Bloomsbury ADVANCES IN THE ANALYSIS OF MICROBIOME DATA Chair: Michele Guindani

E1156: **Training black-box models for de novo reconstruction in metagenomic data**
Presenter: **Sergio Bacallado**, Cambridge University, United States

Metagenomic experiments sequence a mixture of genomes in a sample, for example, a mixture of bacterial genomes from a sample of stool. If there exist reference genomes for the taxa of interest, it is not difficult to assign short reads from a metagenomic dataset to different taxa. When there are taxa which have not been characterised, it becomes necessary to reconstruct their genomes de novo as their abundance in a range of samples is estimated. Current approaches to this problem use our mechanistic understanding of the sequencing process and prior information about the genomes present in the sample. This can involve steps of partial assembly into contiguous sequences, identification of core genes from reference genomes and variable sites within those genes, and statistical modelling of the species abundances in various samples. This process necessarily ignores most of the data, focusing on signals believed reliable a priori. Furthermore, solving the inverse problem at the final step, when species distributions are estimated, can be difficult if uncertainty quantification is desired. We will discuss a different approach, in which we train deep learning models to assign reads to species. The prior information is provided by the training data, which is simulated from a mechanistic model. We explore the reliability of different models, the representations of the data that are learned, and how they are used to classify the reads.

E1268: **Understanding microbiome meta-community structure through a hierarchical Dirichlet process**
Presenter: **Jack OBrien**, Bowdoin College, United States

The explosion of metagenomic sequencing data has led to intense interest in the modeling of microbial community structure that plays an important role in medicine, agriculture, and ecology. We expand on standard mixture model approaches to these data to include a hierarchical Dirichlet process with an underlying latent allocation that allows for more flexible structure within and across samples. We demonstrate an explicit connection between the components of the model (niches) and the unified neutral theory of biodiversity, a key piece of ecological theory. Our formulation allows for a Gibbs sampling strategy and we provide two examples from the recent literature. We also present a thermodynamic integration approach to determining the optimal number of niches in a data set.

E1091: **On current challenges in the analysis of microbiome data from both academic and non-academic perspectives**
Presenter: **Ivo Ugrina**, Faculty of Science, University of Split, Croatia

Microbiome research is increasingly gaining importance in the academic community as well as within the general population due to the enormous potential interactions between human hosts and bacteria might hold towards our well-being. However, as with all promising fields, there are both technical and non-technical difficulties to overcome prior to reaching the full potential of the field. This is especially evident in current approaches to data analysis of microbiome data where a researcher, or an engineer, has to choose among many possible steps in both data preparation and statistical analysis to derive a feasible solution. Some of the challenges to valid statistical conclusions, both in the academic and non-academic communities, will be highlighted and current solutions (and workarounds) will be presented. Special attention will be given to some of the most frequently overlooked aspects of the data analysis pipeline and the hidden dangers accompanying them.

E0880: **Variance component selection with microbiome taxonomic data**
Presenter: **Jin Zhou**, University of Arizona, United States
Co-authors: Jing Zhai, Hua Zhou

High-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. One important problem in microbiome analysis is to identify the bacterial taxa that are associated with a response, where the microbiome data are summarized as the counts or composition of the bacterial taxa at different taxonomic levels. Previous methods consider variable selection of taxonomic data in regression analyses and taxonomic data at different levels are considered as fixed covariates. Due to high-dimensional features of metagenomic information, different penalization schemes have been adopted. On the other hand, the association of microbiome composition and clinical phenotypes was assessed by testing the nullity of variance components, where phylogenetic tree information and distances measures between communities can be incorporated into the model. By combining these two methods we consider regression analysis by treating bacterial taxa at different level as multiple random effects. We propose a variance component selection scheme of high-dimensional taxonomic clusters with Lasso penalization. Our methods seamlessly coupled different distance measures with associated taxonomic selection. Extensive simulations demonstrate the superiority of our methods vs existing methods. Finally, we apply our method to a longitudinal lung microbiome study of HIV infected patients.

EO668 Room Chancellor's Hall TIME SERIES Chair: Andrew Walden

E0396: **Complex oceanic time series observations**
Presenter: **Sofia Olhede**, University College London, United Kingdom

Oceanic data sets take the form of both Eulerian and Lagrangian observations–observations made at fixed spatial points and via drifting instruments. There are challenges in analyzing instruments that are drifting, especially depending on the velocity of drift, as when the instrument moves, the generating parameters change. Traditional time series models can only encapsulate slow variation in the underlying generative mechanism. However, in many scenarios, this is not a realistic assumption. There seems to be an unavoidable conflict between how rapidly the structured part of the model can change, versus how much we need to average in order to retrieve parameters stably. We introduce a new class of nonstationary time series, and show how efficient and rapid inference is still possible in this scenario, despite the generating mechanism changing quickly. The methods are illustrated on drifter time series, from the global drifter programme. Depending on the latitude of the observations, the underlying generative mechanism of the observed phenomenon is either slowly or rapidly changing, and we show how the newly introduced methodology can resolve both scenarios.

E0363: **Time series graphical modelling via partial coherence and lessons from EEG analysis**
Presenter: **Andrew Walden**, Imperial College London, United Kingdom

In several areas of science, such as pollution monitoring and neuroscience, the ability to create a graphical model to help visualise relationships between time series is very useful. One approach to time series graphical modelling utilises partial coherence between pairs of series: it being zero for all frequencies corresponds to a missing edge in the graph. If partial coherence methodology is to be successfully used on EEG data then careful consideration must be given to practical issues such as mixed spectra and poorly condition spectral matrices. The treatment of spectral lines is discussed. Dealing with poor conditioning via spectral matrix shrinkage reveals very different influences on partial coherence estimates may result, depending on the loss function chosen. The production of a meaningful result in practice is much harder than might be anticipated from the theory.

E0461: **A dual parameter long memory time series model**
Presenter: **Rajendra Bhansali**, Imperial College London, United Kingdom
Co-authors: David Natsios
A new model for long-memory time series that characterizes the correlation decay as a mixture of polynomial and logarithmic rates is introduced. This model includes as a special case the standard single-parameter model in which the correlations may decay only at a polynomial rate. Examples illustrating situations where the standard model does not apply, but the new model does do so are presented. A mathematical definition of the class of dual parameter long memory models is given, and extended to include also the class of dual parameter intermediate memory models. The dual parameter fractional ARMA models are also introduced, and the notions of strong, weak and intermediate memory are defined. Non-parametric and semi-parametric estimation of the parameters of this new class of models by suitable extensions of the standard log-periodogram and local Whittle methods is considered, together with the maximum likelihood estimation of the parameters of the dual parameter fractional ARMA model. Asymptotic, and finite sample, properties of the estimates are investigated, and it is shown that the standard single-parameter estimation methods can be badly biased when the dual parameter model applies. The question of discriminating between these two classes of models for observed time series is examined and an application to internet packet traffic is discussed.

E0788: **Complex-valued stochastic process modelling with some physical applications**
Presenter: **Adam Sykulski**, Lancaster University, United Kingdom
In many applications, bivariate time series are represented as complex-valued time series. This representation is useful for separating series that are circular vs noncircular (sometimes referred to as proper vs improper). We present a framework for the parametric modelling of such signals using stochastic processes. We apply our framework to two applications. The first uses a novel widely-linear autoregressive process to model noncircular seismic signals. The second uses a novel anisotropic Matern process to model time series obtained from particle trajectories seeded in fluid dynamic models of turbulence.

EO182 Room Court COPULA MODELING FOR PARTLY DISCRETE DATA OR DATA WITH TIES Chair: Ivan Kojadinovic

E1610: **Inference for copula modeling of discrete data: A cautionary tale and some facts**
Presenter: **Olivier Faugeras**, Toulouse School of Economics - University of Toulouse Capitole, France
Some of the mathematical, statistical and epistemological issues involved in using copulas to model discrete data are elucidated. We contrast the possible use of (nonparametric) copula methods versus the problematic use of parametric copula models. For the latter, we stress, among other issues, the possibility of obtaining impossible models, arising from model misspecification or unidentifiability of the copula parameter.

E0554: **Vine copula regression with mixed discrete and continuous predictors**
Presenter: **Dorota Kurowicka**, Delft University of Technology, Netherlands
The purpose is to show how vine copulas can be used in regression type models where one wants to find the relationship between dependent variable Y and predictive variables $X = (X_1, ..., X_d)$ and see how Y changes for different realizations of the predictive variables. If the relationship between Y and X is linear, the linear regression is a very powerful and efficient for such problem. However, when more complicated relationships between Y and X exist in the data we want to model, it might be beneficial to find conditional distribution of Y given realizations of the predictive variables using vine copulas. This problem has already been discussed in case when Y and X were continuous variables. However, when we allow the variables to be mixed discrete and continuous some difficulties of applying the vine copula regression have to be overcome. The vine copula regression is computationally much more involved than the linear regression but in many cases leads to a better model. We present benefits and challenges of vine copula regression and illustrate its performance on few examples of data sets from different areas of applications.

E0186: **Assessing the structural risk accounting for ties**
Presenter: **Gianfausto Salvadori**, Universita Del Salento, Italy
Co-authors: Fabrizio Durante, Roberta Pappada
Copulas are useful in quite a few different applications, and especially in environmental sciences, where the variables at play are generally non-independent. Usually, these variables are continuous ones, being times, lengths, weights, and so on. Unfortunately, the corresponding observations may suffer from (instrumental) adjustments and truncations, and may show repeated values (i.e., Ties). As a consequence, on the one hand, a tricky issue of model identifiability may arise, and, on the other hand, the assessment of the risk may be adversely affected. A possible remedy consists of suitable data randomization: three different jittering strategies are outlined. A simulation study is carried out in order to evaluate the effects of the randomization of multivariate observations on the risk assessment when ties are present. In particular, it is investigated whether, how, and to what extent, the randomization may change the estimation of the structural risk by using a coastal engineering example, as archetypical of a broad class of models/problems in engineering applications. Practical warnings and advices about the use of randomization techniques are also given.

E0588: **Some copula specification tests adapted to ties**
Presenter: **Ivan Kojadinovic**, CNRS UMR 5142 LMA University of Pau, France
When modeling the distribution of a multivariate continuous random vector using the so-called copula approach, it is not uncommon to have ties in the coordinate samples of the available data because of rounding or lack of measurement precision. Yet, the vast majority of existing inference procedures on the underlying copula were both theoretically derived and practically implemented under the assumption of no ties. Applying them nonetheless can lead to strongly biased results. Some of the existing statistical tests can however be adapted to provide meaningful results in the presence of ties. As it shall be illustrated, this is the case of some tests of exchangeability, radial symmetry, extreme-value dependence and goodness of fit.

EO029 Room G11 RECENT ADVANCES IN HIGH DIMENSIONAL STATISTICS Chair: Ali Shojaie

E1261: **Predictive risk estimation in high-dimensional misspecified quantile regression**
Presenter: **Alexander Giessing**, University of Michigan, Ann Arbor, United States
Predictive modeling plays an important role in many scientific disciplines, including finance and economics. A natural way to gauge predictive models is to estimate their predictive risk. We develop a framework for estimating the predictive risk of high-dimensional quantile regression (QR) models under arbitrary misspecifications. In this framework, the marginal distributions of the response and predictor variables can be non-Gaussian and their relationship can be linear, nonlinear or nonparametric. Our estimator of the predictive risk is based a novel characterization of the down-ward bias of the in-sample risk of the QR fit. In particular, we show that the down-ward bias depends on the specification of the QR model, the covariance of the predictor variables and the conditional density of the response variable. Based on this characterization we propose a de-biased nonparameteric plug-in estimator for the predictive risk and establish its uniform consistency and asymptotic normality. On the theoretical side we provide a new strong Bahadur representation for misspecified QR models with a diverging number of unbounded predictor variables. This new representation fully accounts for any sort of misspecification and therefore retains fast rates of convergence of the remainder term even if the misspecification persists as the sample size increases.

E0909: **Estimating and testing individual mediation effects in high-dimensional settings**
Presenter: **Preetam Nandy**, University of Pennsylvania, United States
Co-authors: Hongzhe Li
The problem of identifying intermediate variables (or mediators) that regulate the effect of a treatment on an outcome is considered. While there has been significant research on this topic, little work has been done when the set of potential mediators is high-dimensional. A further complication arises when the potential mediators are interrelated. In particular, we assume that the causal structure of the treatment, potential mediators and outcome is a directed acyclic graph. In this setting, we propose novel methods for estimating and testing the influence of a mediator on the outcome for high-dimensional linear structural equation models (linear SEMs). For the estimation of individual mediation effect, we propose a modification of the IDA algorithm that was developed for estimating causal effects from observational data. While most of the approaches for estimating the influence of potential mediators ignore the causal relationship among the mediators, our IDA-based approach estimates the underlying causal graph from data. We derive a high-dimensional consistency result for the IDA-based estimators when the data are generated from a linear SEM with sub-Gaussian errors. Further, we propose a first asymptotically valid testing framework in such a setting, leading to a principled FDR control approach for the identification of the set of true mediators. Finally, we empirically demonstrate the importance of using an estimated causal graph in high-dimensional mediation analysis.

E0933: **Double estimation friendly inference in high-dimensional statistics**
Presenter: **Rajen Shah**, University of Cambridge, United Kingdom
Co-authors: Peter Buehlmann, Nicolai Meinshausen
A regression setting is considered where inference concerning the relationship between a response Y and a variable X, after controlling for a high-dimensional vector of additional covariates Z, is of interest. In recent years, great advances have been made in developing methods for this important task, largely based on the debiased Lasso. However, these techniques typically require a sparse linear model of Y on (X,Z) to hold. We will introduce a framework for inference that is valid if either a sparse linear model of Y on Z, or one of X on Z holds. For example, in the latter setting, the methods allow for testing conditional independence of X and Y given high-dimensional covariates Z, and the construction of confidence intervals for the coefficient corresponding to X in a potentially only partially linear model where the contribution of Z to the response is nonlinear. The framework can also be extended to encompass high-dimensional generalised linear models for either the relationship between Y and Z or X and Z.

E1429: **Intelligent sampling for change point analysis problems**
Presenter: **George Michailidis**, University of Florida, United States
Change point estimation is traditionally performed by optimizing over the cases of considering each data point as the true location parameter. In truth, only points close to the actual change point provide useful information for estimation, while data points far away are superfluous, to the point where using only a few points close to the true parameter is just as precise as using the full data set. From this principle we constructed a 2-stage method for change point estimation that localizes the analysis to a few small subsamples of the data but is just as accurate as traditional analysis on the full data set. In fact, we demonstrate that this method achieves the same rate of convergence and even virtually the same asymptotic distribution as the analysis of the full data. Furthermore the subsample of data analyzed can be made so small the entire procedure can run in as low as order of root-N time as opposed to at least $O(N)$ time for all current procedures, making it promising for analysis on long data sets with adequately spaced out change points.

EO045 Room G21A REGULARISATION AND MODEL CHOICE FOR DURATION TIME ANALYSIS Chair: Thomas Kneib

E0643: **Penalized estimation of cumulative effects**
Presenter: **Andreas Bender**, Department of Statistics, LMU Munich, Germany
Modeling effects of time dependent covariates (TDC) introduces additional complexity to time-to-event analysis, especially when these effects are assumed to depend on the complete or partial history of the TDC. Such effects are often referred to as cumulative effects. We present a general framework for penalized estimation of such effects, that potentially vary non-linearly with respect to timing and amount of the TDC as well as over time. Moreover, the time-window of past exposures that may affect the hazard at time t can be specified flexibly. As we embed our approach in the framework of Generalized Additive Mixed Models, robust algorithms for estimation, as well as powerful inference procedures are readily available. We illustrate the proposed method, by investigating the association between daily caloric intake on the intensive care unit and acute survival in critically ill patients.

E1072: **Recent developments on joint modelling for longitudinal and survival data: Applications to biomedicine**
Presenter: **Francisco Gude**, Complexo Hospitalario Universitario de Santiago de Compostela, Spain
Co-authors: Ipek Guler, Christel Faes, Carmen Cadarso Suarez
In many biomedical studies, patients are followed-up repeatedly and several outcomes are recorded in each visit. Clinicians are usually interested in the study of time-to-event with longitudinal biomarkers. Several regression techniques are introduced in the literature to study the association between a longitudinal biomarker and a time-to-event process. Joint modelling approaches are gaining an increasing attention due to their efficiency, reduction in bias and rapidly-expanding literature with software. There is an increasing number of extensions as a result of clinical questions on joint modelling approaches. The aim is to give an overview of recent extensions such as dynamic predictions, multivariate longitudinal case and different association structures. Those extensions will be illustrated with several clinical applications.

E0837: **Variable selection and allocation in joint models for longitudinal and time-to-event data via boosting**
Presenter: **Elisabeth Waldmann**, Friedrich-Alexander-Universitaet Erlangen-Nuernberg, Germany
Co-authors: Andreas Mayr, Colin Griesbach
Joint Models for longitudinal and time-to-event data have gained a lot of attention in the last few years as they are a helpful technique to approach a data structure common in clinical studies where longitudinal outcomes are recorded alongside event times. Those two processes are often linked and the two outcomes should thus be modeled jointly in order to prevent the potential bias introduced by independent modelling. Commonly, joint models are estimated in likelihood based expectation maximization or Bayesian approaches using frameworks where variable selection is problematic and which do not immediately work for high-dimensional data. A boosting algorithm rendered possible the selection of covariates even in high-dimensional settings. This algorithm is extended to the automated variable allocation. This approach is necessary when there is no prior knowledge available on the question which dependent variable the individual covariates have influence on.

E0201: **Selection of effects in Cox frailty models by regularization methods**
Presenter: **Andreas Groll**, Georg-August-University Gottingen, Germany
Co-authors: Thomas Kneib, Trevor Hastie, Gerhard Tutz
In all sorts of regression problems it has become more and more important to deal with high dimensional data with lots of potentially influential covariates. A possible solution is to apply estimation methods that aim at the detection of the relevant effect structure by using regularization methods. The effect structure in the Cox frailty model, which is the most widely used model that accounts for heterogeneity in survival data, is

investigated. Since in survival models one has to account for possible variation of the effect strength over time the selection of the relevant features has to distinguish between several cases, covariates can have time-varying effects, can have time-constant effects or be irrelevant. A regularization approach is proposed that is able to distinguish between these types of effects to obtain a sparse representation that includes the relevant effects in a proper form. The method is applied to a real world data set, illustrating that the complexity of the influence structure can be strongly reduced by using the proposed regularization approach.

EO461 Room G3 MODEL SELECTION AND INFERENCE Chair: Ulrike Schneider

E1120: **Valid tests for sufficient dimension reduction**
Presenter: **Kory Johnson**, University of Vienna, Austria
Many testing situations produce a table of p-values which are interpreted sequentially to conduct inference. Most notably, stepwise regression tests for significant reduction in error sum of squares at every step. While these p-values are unadjusted for selection, a method has been recently provided for validly using such p-values to determine the stepwise model. We present extensions to demonstrate its generality. In particular, we consider sufficient dimension reduction, which produces similar p-value tables. As in the stepwise case, these p-values do not jointly represent valid tests of the corresponding hypotheses, because data-based selection is performed prior to their construction. By appropriate adjustments of the rejection thresholds, however, these unadjusted p-values can be used to validly select the appropriate dimension. These adjusted thresholds are motivated by alpha-investing, which provides a new framework for thinking about sequential, multiple comparison problems.

E0992: **Prediction when fitting simple models to high-dimensional data**
Presenter: **Hannes Leeb**, University of Vienna, Austria
Co-authors: Lukas Steinberger
Linear subset regression is studied in the context of a high-dimensional linear model. Consider $y = \vartheta + \theta' z + \varepsilon$ with univariate response y and a d-vector of random regressors z, and a submodel where y is regressed on a set of p explanatory variables that are given by $x = M'z$, for some $d \times p$ matrix M. Here, 'high-dimensional' means that the number d of available explanatory variables in the overall model is much larger than the number p of variables in the submodel. We present Pinsker-type results for prediction of y given x. In particular, we show that the mean squared prediction error of the best linear predictor of y given x is close to the mean squared prediction error of the corresponding Bayes predictor $E[y|x]$, provided only that $p/\log d$ is small. We also show that the mean squared prediction error of the (feasible) least-squares predictor computed from n independent observations of (y,x) is close to that of the Bayes predictor, provided only that both $p/\log d$ and p/n are small. Our results hold uniformly in the regression parameters and over large collections of distributions for the design variables z.

E0998: **Hypothesis testing when fitting simple models to high-dimensional data**
Presenter: **Lukas Steinberger**, University of Freiburg, Germany
Co-authors: Hannes Leeb
In a linear regression problem with a huge number of potentially important explanatory variables, it is often desirable to select only a few regressors, even if it is not a priori evident that most regressors are truly irrelevant or at least practically negligible. If such a simple working model is maintained, even though the true data generating mechanism is much more complex, the subset regression problem may not be linear and homoskedastic, due to the omission of important variables, even if the full model is. Therefore, classical linear regression techniques may be inappropriate for the working model. In contrast to these well known issues, we show that if the number of available explanatory variables in the full model is very large, then the classical f-test, based on the observations of only a few regressors of interest, is approximately valid for testing the significance of these regressors for prediction. Therefore, we conclude that although a small set of regressors with high predictive power may not even exist, or it may be computationally prohibitive to find, it is still possible to decide about the predictive power of any given choice of regressors.

E0628: **Uniformly valid confidence intervals post-model-selection**
Presenter: **David Preinerstorfer**, Universita libre de Bruxelles, Belgium
Co-authors: Francois Bachoc, Lukas Steinberger
General methods are suggested to construct asymptotically uniformly valid confidence intervals post-model-selection. The constructions are based on principles recently proposed in the literature. In particular, the candidate models used can be misspecified, the target of inference is model-specific, and coverage is guaranteed for any data-driven model selection procedure. After developing a general theory, we apply our methods to practically important situations where the candidate set of models, from which a working model is selected, consists of fixed design homoskedastic or heteroskedastic linear models, or of binary regression models with general link functions.

EO636 Room G4 REPURPOSING DRUGS USING ELECTRONIC HEALTH RECORDS Chair: Roy Welsch

E1253: **Statistical considerations for synthetic clinical trials**
Presenter: **Rebecca Betensky**, Harvard School of Public Health, United States
Electronic health records constitute a rich resource for studying many aspects of disease, including effects of drug exposures. It is tempting to use these data for discovery of unexpected, effective drugs for given conditions. We will review the literature on this emerging research area and evaluate this approach with regard to considerations of statistical inference. We will discuss an example in Alzheimer's disease.

E1192: **Studentized sensitivity analysis in paired observational studies**
Presenter: **Colin Fogarty**, Massachusetts Institute of Technology, United States
A fundamental limitation of causal inference in nonrandomized studies, including observational studies, broken experiments, and synthetic clinical trials, is that perceived evidence for an effect may well be explained away by factors not accounted for in the primary analysis. This deficiency necessitates an additional step, known as a sensitivity analysis, which assesses how strong unmeasured confounding would have to be in order to materially alter the findings of the study. Methods for conducting a sensitivity analysis have been established under the assumption that the treatment effect is constant across all individuals. This assumption is viewed as innocuous in the analysis of randomized experiments, as the variance for the estimated average treatment effect is maximized if the treatment effect is additive; however, in the potential presence of unmeasured confounding, it has been argued that certain patterns of effect heterogeneity may render the performed sensitivity analysis inadequate. We present a new method for conducting a sensitivity analysis in the presence of heterogeneous treatment effects. The method naturally extends conventional tests for the sample average treatment effect in a randomized experiment to the case of unknown, but bounded, probabilities of assignment to treatment. In so doing, we illustrate that concerns about the restrictiveness of the constant treatment effect model are largely unwarranted.

E1499: **Investigating anticipated and un-anticipated effects of drugs in large biobanks**
Presenter: **Ioanna Tzoulaki**, Imperial College London, United Kingdom
Large biobank studies with linked electronic health records (EHR) offer unprecedented opportunities to study the effects of drugs in a systematic way in order to uncover anticipated and anticipated effects associated with drug use. We will present two examples of using multidimensional longitudinal EHR data to examine repurposing drugs in a prospective design. On one instance, a cohort study using the Clinical Practice Research

Datalink was designed to investigate the association between use of metformin compared with other antidiabetes medications and cancer risk by emulating an intention-to-treat analysis as in a trial. Cox proportional hazards models were used to estimate multivariable-adjusted hazard ratios (HR) and 95% CI. In another example, we use human genetic variation as a proxy and a method of more accurately predicting the physiologic effects of drugs. Specifically, we apply a systematic approach to examine a wide range of phenotypes across the human phenome in relation to genetic variants mimicking the effects of drugs of interest (Phenome Wide Association Study). This approach is now tractable given the extensive phenotypic information available in large biobanks (e.g.UK Biobank) with linkage to electronic health records and measurements of quantitative phenotypes at baseline. We present an example based on genetic variants on the HMGR gene which mimics the effect of statins and show the phenome wide association study related to those variants.

E1841: **Repurposing evidentiary standards and evaluation in practice**
Presenter: **Virginia Acha**, MSD, United Kingdom
Co-authors: Adrian Towse
The aim is to address the challenges of defining repurposing and the resulting confusion in evidentiary requirements and regulatory review to support use for patients in practice. These challenges have been made more acute by the social conflicts in incentives and understanding, resulting from wider scope of stakeholders directly engaging in establishing what appropriate evidentiary standards and evaluation should be and where responsibility and accountability for these should lie. We use recent examples from the UK to illustrate these points, and conclude with proposals for new research to address the gaps.

EO164 Room G5 COMPUTATIONAL ASPECTS IN INFERENCE FOR STOCHASTIC PROCESSES Chair: Hiroki Masuda

E1089: **Curse of dimensionality of MCMC**
Presenter: **Kengo Kamatani**, Osaka University, Japan
Some Markov-chain Monte Carlo methods for complex target distributions are studied and reviewed. We consider the effect of dimensionality along with the tail property of the target distribution. We also apply the results to the Bayesian inference for stochastic processes.

E0679: **New classes and methods in YUIMA**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
Some advances in the implementation of advanced mathematical tools in the Yuima package are presented. First, we discuss a new class called yuima.law that refers to the mathematical description of a general Levy process used in the formal definition of a general Stochastic Differential Equation. The final aim is to have an object, defined by the user, that contains all possible information about the Levy process considered. This class creates a link between Yuima and other R packages that manage specific Levy process like for example ghyp or MixedTS available on CRAN. The second class, that we discuss, is the yuima.PPR that refers to the mathematical description of the Point Process Regression Model. This model can be seen as a generalization of a self-exciting point process, since it is possible to consider external covariates in the intensity process. We discuss the implemented methods for the definition, simulation and estimation of these processes.We also discuss the exact maximum likelihood estimation procedure for a Levy CARMA model. The new method is based on the existence of an analytical expression for the characteristic function of a general Levy CARMA model. We show that we are able to overcome the problems that arise in the estimation procedure based on Kalman Filter that requires the existence of the second moment for the underlying Levy process. Real and simulated examples are shown.

E0370: **Intensities ratios models with applications to the modelling of limit order books**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
A Cox-type model is introduced for relative intensities of orders flows in a limit order book. The Cox-like intensities of the counting processes of events are assumed to share an unobserved and unspecified baseline intensity, which in finance can be identified to a global market activity affecting all events. The model is formulated in terms of relative responses of the intensities to covariates, and relative parameters can be estimated via quasi likelihood maximization. Consistency and asymptotic normality of the estimators are proven. Computationally intensive inferences are run on large samples of tick-by-tick data (35+ stocks and 220+ trading days, adding to more than one billion events). Results of the model are interpreted in terms of probability of occurrence of events. Excellent agreement with empirical data is found. Estimated model reproduces known empirical facts on imbalance, spread and queue sizes, and helps identifying trading signals of interests on a given stock.

E0974: **Penalized methods for stochastic processes**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan
Quasi likelihood analysis (QLA) is a general framework for statistics for stochastic processes. The QLA established asymptotic (mixed) normality and convergence of moments of the quasi maximum likelihood estimator and the quasi Bayesian estimator for various dependency models such as diffusion processes, even under discrete sampling, and point processes. Recently applications of the QLA have been extending to regularization methods for sparse estimation. The L_q penalized quasi likelihood function methods and the least squares approximation methods are discussed as well as their applications to volatility estimation for diffusion processes and a regression analysis for point processes.

EO559 Room Gordon ADVANCES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Hao Chen

E0367: **A general framework for information pooling in two-sample sparse inference**
Presenter: **Yin Xia**, Fudan University, China
A general framework is developed for exploiting the sparsity information in two-sample multiple testing problems. We propose to first construct a covariate sequence, in addition to the usual primary test statistics, to capture the sparsity structure, and then incorporate the auxiliary covariates in inference via a three-step algorithm consisting of grouping, adjusting and pooling (GAP). The GAP procedure provides a simple and effective framework for information pooling. An important advantage of GAP is its capability of handling various dependence structures such as those arise from multiple testing for high-dimensional linear regression, differential correlation analysis, and differential network analysis. We establish general conditions under which GAP is asymptotically valid for false discovery rate control, and show that these conditions are fulfilled in a range of applications. Numerical results demonstrate that existing methods can be significantly improved by the proposed framework. The GAP procedure is illustrated using a breast cancer study for identifying gene-by-gene interactions.

E0803: **Memory efficient low-rank approximation from incomplete entries via nonconvex optimization**
Presenter: **Xiaodong Li**, UC Davis, United States
Co-authors: Ji Chen
Low-rank approximation is one of the most useful tools in statistics and machine learning. It is used for dimension reduction such as PCA, ICA, CCA, kernel PCA, etc. It is also used for memory efficient computing such as kernel approximation. Due to large sample sizes in machine learning applications, standard low-rank approximation algorithms, such as singular value decomposition, usually cannot be directly applied. Taking kernel PCA for an example, it is intensive in both memory and computation to obtain and store the whole Gram matrix. Inspired by the recent development of nonconvex matrix completion, we propose a general framework of rank-r approximation from a few entries, without any spectral or incoherence

assumptions on the underlying matrix. Theoretically, we will present a general result whose corollaries improve or match state-of-the-art results for different setups of matrix completion in the literature. Moreover, we apply our method to randomized kernel PCA to showcase its effectiveness and usefulness.

E0639: **Maximum likelihood inference for a large precision matrix**
Presenter: **Yunzhang Zhu**, Ohio State University, United States
Co-authors: Xiaotong Shen, Wei Pan
Inference concerning Gaussian graphical models involves pairwise conditional dependencies on Gaussian random variables. In such a situation, regularization of a certain form is often employed to treat an overparameterized model, imposing challenges to inference. We will present a constrained maximum likelihood method for inference, with a focus of alleviating the impact of regularization on inference. For general composite hypotheses, we unregularize hypothesized parameters whereas regularizing nuisance parameters through a L_0-constraint controlling the degree of sparseness. For the likelihood ratio test, the corresponding distribution is the chi-square or normal, depending on if the co-dimension of a test is finite or increases with the sample size. This goes beyond the classical Wilks phenomenon. Some numerical results will be discussed, in additional to an application to network analysis.

E0920: **A unified nonparametric procedure on detecting spurious discoveries under sparse signals**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Chao Zheng, Wenxin Zhou, Lyuou Zhang
Identifying a subset of response-associated covariates from a large number of candidates has become a fundamental tool for scientific discoveries, particularly in biology including the differential analysis in genomics and the genome-wide association study in genetics. However, given the high dimensionality and the sparsity of signals in data, spurious discoveries can easily arise. Also, the ubiquitous data with mixed types, along with complex dependence, greatly limit the applicability of the traditional goodness-of-fit-based procedures. We introduce a statistical measure on the goodness of spurious fit based on the maximum rank correlations among predictors and responses. The proposed statistic imposes no assumptions on the data types, dependency, and the underlying models. We derive the asymptotic distribution of such goodness of spurious fit under very mild assumptions on the associations among predictors and responses. Such an asymptotic distribution depends on the sample size, ambient dimension, the number of predictors under study, and the covariance information. We propose a multiplier bootstrap procedure to estimate such a distribution and utilize it as the benchmark to guard against spurious discoveries. It is also applied to the variable selection problems for the high dimensional generalized regressions. We applied our method to genetic studies to demonstrate that the proposed measure provides a statistical verification of the detected biomarkers.

EO206 Room CLO 101 HIGHER MOMENTS IN MULTIVARIATE ANALYSIS Chair: Nicola Loperfido

E0308: **Maximal skewness projections for scale mixtures of skew-normal vectors**
Presenter: **Jorge Martin Arevalillo**, UNED, Spain
Co-authors: Hilario Navarro Veguillas
Multivariate scale mixtures of skew-normal (SMSN) variables are flexible models that account for non-normality in multivariate data scenarios by tail weight assessment and a shape vector representing the asymmetry of the model in a directional fashion. Its stochastic representation involves a skew-normal (SN) vector and a non negative mixing scalar variable, independent of the SN vector, that injects kurtosis into the SMSN model. We address the problem of finding the maximal skewness projection for vectors that follow a SMSN distribution; when simple conditions on the moments of the mixing variable are fulfilled, it can be shown that the direction yielding the maximal skewness is proportional to the shape vector. This finding stresses the directional nature of the asymmetry in this class of distributions; it also provides the theoretical underpinnings for skewness based projection pursuit under a general and flexible class of multivariate distributions. Some examples that show the validity of our result for some widely used distributions within the SMSN family are also given.

E0797: **Fourth cumulant for multivariate aggregate claim models**
Presenter: **Farrukh Javed**, Orebro University, Sweden
Co-authors: Nicola Loperfido, Stepan Mazur
The fourth cumulant for the aggregated multivariate claims is considered. We present a formula for the general case when the aggregating variable is independent of the multivariate claims. Two important special cases are also considered. In the first one, multivariate skewed normal claims are considered and aggregated by a Poisson variable. The second case is dealing with multivariate asymmetric generalized Laplace and aggregation is made by a negative binomial variable. Due to the invariance property, the latter case can be derived directly, leading to the identity involving the cumulant of the claims and the aggregated claims. There is a well established relation between asymmetric Laplace motion and negative binomial process that corresponds to the invariance principle of the aggregating claims for the generalized asymmetric Laplace distribution. We explore this relationship and provide multivariate continuous time version of the results. It is discussed how these results that deal only with dependence in the claim sizes can be used to obtain a formula for the fourth cumulant for more complex aggregate models of multivariate claims in which the dependence is also in the aggregating variables.

E0665: **Linear projections for kurtosis removal**
Presenter: **Nicola Loperfido**, University of Urbino, Italy
The performance of several statistical procedures might depend on the kurtosis of the sampled distribution. Examples include, but are not limited to, inference on mean vectors and covariance matrices. Componentwise, nonlinear transformations might not adequately address the problem. We propose to remove kurtosis by means of linear projections which are solutions of generalized eigenvalue problem involving two kurtosis matrices. The method is particularly useful for either hidden truncation models or finite mixture models. Its practical usefulness is illustrates with the RANDU dataset.

E1384: **Gamma distributed covariance matrices and their moments**
Presenter: **Krzysztof Podgorski**, Lund University, Sweden
Co-authors: Stepan Mazur, Tomasz Kozubowski
The class of matrix gamma distributions with fixed matrix parameters is closed on convolutions with respect to the shape parameter. The convolution property extends to the singular Wishart distributions. However, the gamma matrix distribution is not infinitely divisible in the usual sense as pointed, previously. The problem and its solution has a long history, probably initiated by Levy in 1948. Later, infinite divisibility of the Wishart distribution (or the lack thereof), and its entries, has been discussed in detail on many occasions. We extend the concept of the matrix value gamma distributions to the singular case and discuss their properties. We give a natural stochastic representation of this family of matrix value distributions parametrized by a non-negative time parameter. The group property with respect to this positive time-like parameter is discussed. We show how this property can be utilized to construct continues time models using random gamma matrices despite the lack of infinite-divisibility. We present moment properties of the introduced models.

EO360 Room CLO 102 INFERENCE IN SEMI- AND NONPARAMETRIC MODELS Chair: Christopher Parmeter

E1069: **Accelerated failure time models with logconcave errors**
Presenter: **Ruixuan Liu**, Emory University, United States
Accelerated failure time models are studied where the survivor function of the error term is logconcave. The logconcavity assumption is natural to represent the ageing or wear-out effect of the failure time and it is often assumed in related economic models. We construct semiparametric maximum likelihood estimates subject to the shape restriction with right censored data. We establish large sample properties of the resulting estimates for the finite dimensional parameter and show our estimators achieve the semiparametric efficiency bound. Our shaped restricted estimation also facilitates the computation, as in the proposed iterative procedure it leads to monotone estimating equations in each iteration step.

E1000: **Unified estimation of densities on bounded and unbounded domains**
Presenter: **Carlos Martins-Filho**, University of Colorado at Boulder, United States
Kernel density estimation in domains with boundaries is known to suffer from undesirable boundary effects. We show that in the case of smooth densities, a general and elegant approach is to estimate an extension of the density. The resulting estimators in domains with boundaries have biases and variances expressed in terms of density extensions and extension parameters. The result is that they have the same rates at boundary and interior points of the domain. Contrary to the extant literature, our estimators require no kernel modification near the boundary and kernels commonly used for estimation on the real line can be applied. Densities defined on the half-axis and in a unit interval are considered. The results are applied to estimation of densities that are discontinuous or have discontinuous derivatives, where they yield the same rates of convergence as for smooth densities on R.

E1075: **Departures from symmetry and testing for heteroskedasticity in nonparametric regression**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Alice Sheehan
A new conditional moment test for heteroskedasticity in nonparametric regression models is proposed. The test statistic uses kernel based estimation. Our test builds on previous work, but relaxes the assumption that the density function have compact support. This eliminates the need to trim small values of the data to ensure the test statistic is well behaved. We then introduce a naive bootstrap method as an alternative to the wild bootstrap method previously employed. We show asymptotic equivalence of our test statistic and bootstrap method. A Monte Carlo simulation is provided, with critical values obtained from both the wild bootstrap and a naive bootstrap, to assess the finite sample performance of the proposed test as well as the bootstrap method.

E0984: **Calculating degrees of freedom in multivariate local polynomial regression**
Presenter: **Christopher Parmeter**, University of Miami, United States
Co-authors: Nadine McCloud
The matrix that transforms the response variable in a regression to its predicted value is commonly referred to as the hat matrix. The trace of the hat matrix is a standard metric for calculating degrees of freedom. Nonparametric-based hat matrices do not enjoy all properties of their parametric counterpart in part owing to the fact that the former do not always stem directly from a traditional ANOVA decomposition. In the multivariate, local polynomial setup with a mix of continuous and discrete covariates, which include some irrelevant covariates, we formulate asymptotic expressions for the trace of the resultant hat matrix from the estimator of the unknown conditional mean and derivatives, as well as asymptotic expressions for the trace of the ANOVA-based hat matrix from the estimator of the unknown conditional mean. For a bivariate regression, we show that the asymptotic expression of the trace of the former hat matrix associated with the conditional mean estimator is equal up to a linear combination of kernel-dependent constants to that of the ANOVA-based hat matrix. Additionally, we document that the ANOVA-based hat matrix converges to 0 in any setting where the bandwidths diverge. This can occur in the presence of irrelevant continuous covariates or it can arise when the underlying data generating process is in fact of polynomial order. We use simulated examples to demonstrate that our aforementioned theoretical contributions are valid in finite-sample settings.

EO354 Room Jessel PROVABLE TENSOR METHODS IN MACHINE LEARNING Chair: Will Wei Sun

E0284: **Tensor canonical correlation analysis**
Presenter: **Eric Chi**, North Carolina State University, United States
Canonical correlation analysis (CCA) is a popular multivariate analysis technique that finds the linear relationship between two data sets. Recent technologies such as neuroimaging and remote sensing generate data in form of multi-dimensional arrays or tensors. Classic CCA is insufficient for dealing with tensor data due to the multi-dimensional structure and ultra high-dimensionality. We present tensor CCA, a technique that discovers linear relationship between two tensor data sets while respecting the spatial information. We delineate various population models and propose efficient and scalable estimation algorithms that have global convergence guarantees. Simulation studies illustrate the performance of our method.

E0371: **Covariate-adjusted tensor classification in high-dimensions**
Presenter: **Xin Zhang**, Florida State University, United States
In contemporary scientific research, it is of great interest to predict a categorical response based on a high-dimensional tensor and additional covariates. We introduce the CATCH model (in short for Covariate-Adjusted Tensor Classification in High-dimensions), that efficiently integrates the covariates and the tensor to predict the categorical outcome and jointly explains the relationships among the covariates, the tensor predictor, and the categorical response. To tackle the new computational and statistical challenges arising from the intimidating tensor dimensions, we propose a group penalized approach and an efficient algorithm. Theoretical results confirm that our method achieves variable selection consistency and optimal prediction, even when the tensor dimension is much larger than the sample size. The superior performance of our method over existing methods is demonstrated in extensive simulation studies, a colorimetric sensor array data, and two neuroimaging studies.

E0975: **Spacey random walks**
Presenter: **Austin Benson**, Cornell University, United States
Co-authors: David Gleich, Lek-Heng Lim
Random walks are a fundamental model in applied mathematics and are a common example of a Markov chain. A standard way to compute the stationary distribution for a random walk on a finite set of states is to compute the Perron vector of the associated transition probability matrix. There are algebraic analogues of the Perron vector in terms of z eigenvectors of transition probability tensors whose entries come from higher-order Markov chains. These vectors look stochastic, but they are derived from an algebraic substitution in the stationary distribution equation of higher-order Markov chains and do not carry a probabilistic interpretation. We present the spacey random walk, a non-Markovian stochastic process whose stationary distribution is given by a dominant z eigenvector of the transition probability tensor. The process itself is a vertex-reinforced random walk, and its discrete dynamics are related to a continuous dynamical system. We analyze the convergence properties of these dynamics and discuss numerical methods for computing the stationary distribution. We also provide several applications of the spacey random walk model in population genetics, ranking, and clustering data, and we use the process to analyze taxi trajectory data in New York.

E1161: **Estimation accuracy and computational efficiency of non-parametric kernel tensor estimators**
Presenter: **Taiji Suzuki**, University of Tokyo, Japan

A problem of estimating a low rank tensor in infinite dimensional Hilbert spaces is considered, and the statistical properties and computational efficiency of some estimators is discussed. There are wide applications of estimating low rank tensors, for example, recommendation system, spatio-temporal data analysis, and multi-task learning. The tensor model we consider consists of the sum of products of functions in Reproducing Kernel Hilbert Spaces (RKHSs) defined on the input data sources. To estimate the nonparametric model, we consider a Bayes estimator and an empirical risk minimizer computed by an alternating minimization method. Then, we discuss the trade-off between the statistical performances and the computational efficiency by showing some theoretical and numerical results. In the theoretical analysis, we give upper bounds of the risks and show the minimax optimal rate.

EO717 Room Montague MODEL SELECTION AND CLASSIFICATION FOR VERY LARGE SCALE DATA Chair: Maria-Pia Victoria-Feser

E1283: **On an exact nonparametric test for class separability for the purpose of filter-type model selection**
Presenter: **Fabian Schroeder**, AIT Austrian Institute of Technology GmbH, Austria
Co-authors: Peter Filzmoser

The aim is to introduce and to discuss a statistical test based on the prediction error of a nonparametric classifier for the purpose of filter-type variable selection. This approach has several advantages. First, it naturally accounts for the operating conditions of the classification task, comprising the misclassification costs of the class distribution in the overall population. Many common filter statistics, e.g. the t-test disregard the operation conditions as well as differences in the variances between the class conditionals, which may lead to false conclusions. Secondly, the nonparametric approach guarantees that the differences in the test statistic do not only reflect the differences in the extent to which the variables satisfy the distributional assumption. It is, thus, robust with respect to deviations from the distributional assumptions, e.g. skewness or outliers. Thirdly, it is possible to obtain the exact finite sample distribution of the test statistic under the assumption of equal class conditionals. The exact significance of the classification result, as given by the p-value, is very handy since it gives an absolute criterion for the filter rather than just a relative one, which only allows ranking the variables. Furthermore it also allows adjusting the p-values for the number of tests performed.

E0908: **A prediction divergence criterion for model selection in high dimensional settings**
Presenter: **Maria-Pia Victoria-Feser**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Marco Avella Medina

A new class of model selection criteria is proposed which is suited for stepwise approaches or can be used as selection criteria in penalized estimation based methods. This new class, called the d-class of error measure, generalizes Efron's q-class. This class not only contains classical criteria such as Mallow's Cp or the AIC, but also enables one to define new criteria that are more general. Within this new class, we propose a model selection criterion based on a prediction divergence between two nested models' predictions that we call the Prediction Divergence Criterion (PDC). The PDC provides a different measure of prediction error than a criterion associated to each potential model within a sequence and for which the selection decision is based on the sign of differences between the criteria. The PDC directly measures the prediction error divergence between two nested models and provides different criteria. As an example, we consider linear regression models and propose a PDC criterion that is the direct counterpart of Mallow's Cp. We show that a selection procedure based on the PDC, compared to the Cp, has a smaller probability of overfitting and a negligible asymptotic probability of selecting a larger model for models with more than one additional non significant covariate. The PDC is particularly well suited in high dimensional and sparse situations and also under small model misspecifications.

E1273: **A scalable frequentist model averaging method**
Presenter: **HaiYing Wang**, University of Connecticut, United States

Frequentist model averaging is an effective technique to handle model uncertainty. However, calculation of the weights for averaging is extremely difficult, if not impossible, even when the dimension of the predictor vector, p, is moderate, because we may have 2^p candidate models. We propose a scalable frequentist model averaging method to overcome this difficulty by using the singular value decomposition. The method enables us only need to find the optimal weights for at most p candidate models. We prove that the minimum loss of the scalable model averaging estimator is asymptotically equal to that of the traditional model averaging estimator, and that the scalable Mallows/Jackknife model averaging estimators are asymptotically optimal. We also further extend the method for the high-dimensional case (i.e., $p \geq n$). Numerical studies illustrate the superiority of the proposed method in terms of both the statistical efficiency and computational cost.

E1374: **A prediction-based algorithm for variable selection with applications in genomics**
Presenter: **Roberto Molinari**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Yanyuan Ma, Marco Avella Medina, Samuel Orso, Mili Nabil

The task of model selection is often associated with the minimization of a given loss function which, in the vast majority of cases, is linked to the objective function used to estimate the model parameters (e.g. the likelihood function). However, in many applied cases, these loss functions are not necessarily what practitioners are interested in minimizing. For this purpose, we propose a new algorithm which, among others, makes use of cross-validation to deliver tailor-made variable selection criteria which respond to the needs of practitioners. This approach is flexible in terms of the modelling framework of reference and in terms of criteria of interest as well as being able to deliver a set of sparse models with extremely high predictive power. The latter is particularly useful in the field of genomics where, among millions of gene transcripts, it is not only important to select a few genes which, for example, can predict the presence of certain diseases but can also build gene networks that can have important biological interpretations. Some applied examples show how the proposed method not only performs better than existing approaches in terms of prediction accuracy but also delivers a flexible framework that opens new avenues for model estimation and variable selection in large data settings.

EO258 Room Senate ROBUSTNESS FOR HIGH-DIMENSIONAL AND FUNCTIONAL DATA Chair: Stefan Van Aelst

E0543: **Fast and robust model selection criterion in generalized linear models**
Presenter: **Kris Peremans**, University of Leuven, Belgium
Co-authors: Stefan Van Aelst

Selecting the optimal model from a set of competing models is an essential task in statistics. The focus is on selecting the best subset of available explanatory variables in generalized linear models. It is well-known that standard model selection criteria for generalized linear models are extremely sensitive to contamination in the data. Therefore, robust alternatives are introduced. Particular attention is paid to robust model selection criteria based on resampling techniques. However, a recalculation of robust criteria for each resample is computer intensive because robust criteria are already computationally intensive compared to their non-robust versions. To reduce the computational burden, a modified resampling procedure, inspired by the fast and robust bootstrap method, is proposed. Moreover, it is shown that this modification still yields consistent model selection criterion, in the sense that the optimal model is selected with probability one as the sample size grows to infinity. The performance of the proposed methodology is evaluated empirically by a simulation study and illustrated on real data examples.

E1300: **Centrality measures for graphons**
Presenter: **Marco Avella Medina**, MIT, United States
Co-authors: Francesca Parise, Michael Schaub, Santiago Segarra
Networks and graphs have become prevalent tools to represent systems across various scientific domains, as they provide a natural way to model and explain systems with pairwise interactions. In particular, node centrality measures have become of foremost importance in graph analysis for their ability to identify the most relevant agents and topological features in a network. The size of the networks under analysis has constantly increased in the last years and, consequently, graphons have been introduced as the natural limiting objects for graphs of increasing size. Formal definitions of centrality measures for graphons are introduced and their connections to centrality measures defined on finite graphs are established. In particular, we build on the theory of linear integral operators to define degree, eigenvector, and Katz centrality functions for graphons. We then establish concentration inequalities showing that these centrality functions are natural limits of their analogous counterparts defined on a sequence of random graphs of increasing size. We discuss how to compute our centrality measures as well as their stability towards noise. Our findings are illustrated through a set of detailed numerical examples.

E1401: **Fast robust correlations with application to cellwise outliers**
Presenter: **Peter Rousseeuw**, KU Leuven, Belgium
Co-authors: Jakob Raymaekers
The Pearson (product-moment) correlation coefficient is sensitive to outliers in the data. The literature contains quite a few other correlation measures which are more robust to outliers. Many of these have a higher computational complexity than Pearson's O(n), especially if they are computed from very robust scatter matrices. But in high-dimensional data there are many variables to correlate, so speed is essential. We compare several robust correlation measures with a view toward computation time, preferably no more than O(n log(n)) for a pair of variables, and the ability to form positive semidefinite correlation matrices. We then apply the selected measures to multivariate data analysis techniques including cluster analysis, classification and the detection of cellwise outliers.

E1455: **Performance analysis and robustness for dimension reduction**
Presenter: **Shojaeddin Chenouri**, University of Waterloo, Canada
Information in the data often has far fewer degrees of freedom than the number of variables encoding the data. Dimensionality reduction attempts to reduce the number of variables used to describe the data. There are several dimensionality reduction methods available in the literature for linear and nonlinear manifolds. Each method works only under certain underlying assumptions. There is no universal agreement on how to assess and compare the performance of these different methods, and their robustness properties have not been studied. We attempt to discuss these issues and provide some answers. We introduce a goodness measure called local Spearman correlation for assessing performance and then employ it to define types of influence function and breakdown point to study the robustness of dimensionality reduction methods.

EO174 Room Woburn RECENT AVANCES IN NETWORK MODELING Chair: Veronica Vinciotti

E0549: **Estimating the number of clusters in a stochastic block model for temporal networks**
Presenter: **Riccardo Rastelli**, Vienna University of Economics and Business, Austria
Co-authors: Pierre Latouche, Nial Friel
The latent Stochastic Block Model is a flexible and widely used statistical model for networks that can be generalized to a dynamic context, where interactions are observed at a number of different time frames. This model hypothesizes a latent clustering structure on the nodes of the network, and for this reason it shares many similarities with mixture models. In this context, techniques such as the Expectation-Maximization algorithm usually require a variational approximation on the posterior distribution to be used. We will show that, under general assumptions, it is possible to analytically integrate out most of the model parameters of a dynamic Stochastic Block Model. This leads to an exact formula for a model-based criterion: the Integrated Completed Likelihood (ICL). In this formulation the ICL depends on the model parameters only through the nodes' allocations. We will illustrate a greedy discrete optimization tool that can maximize the ICL, hence returning both the optimal clustering and the optimal model in one algorithmic framework. One important advantage of this approach is that it does not rely on any approximation of the posterior distribution or of the model-based criterion. The algorithm proposed over-performs other existing procedures both in terms of model-choice accuracy and computational complexity. We will show applications of the methodology to both simulated networks and to a real dataset of bike hires in London.

E1153: **Preferential attachment and arrival times**
Presenter: **Benjamin Bloem-Reddy**, University of Oxford, United Kingdom
Well-known models such as the Chinese Restaurant process (CRP), the Yule-Simon process, and the Preferential Attachment model are constructed from a sequence of size-biased sampling steps. Interest in these models is due in large part to the fact that power law degree distributions emerge from a simple generative mechanism. Starting with the observation that these and other models differ only in the distribution of arrival times, a class of random graph models called (α,t)-graphs is defined. (α,t)-graphs admit a representation in terms of products of independent beta random variables reminiscent of the so-called stick-breaking representation of the CRP, and their limiting degree sequences give rise to a number of distributional identities. Posterior inference is straightforward, and a Gibbs sampler is used to fit some example models to data.

E0734: **Credit risk modelling via efficient network-based inference**
Presenter: **Veronica Vinciotti**, Brunel University London, United Kingdom
Co-authors: Elisa Tosetti, Francesco Moscone
A credit risk model with network effects for a very large number of companies is developed. We assume a probit specification with group-specific effects having a non-diagonal, sparse covariance matrix, and adopt a penalised maximum likelihood estimation approach. Hence, we develop an Expectation-Maximization algorithm where we exploit the properties of truncated normals to proxy the conditional expectations. Monte Carlo experiments show that our proposed estimator has good finite sample properties and can be adopted for estimation and prediction using very large, or huge, datasets, given its moderate computation costs.We use a sample of nearly 568,000 accounts for Small and Medium-sized Enterprises (SMEs) in the United Kingdom in the years 2009 to 2013. We compare the prediction performance and estimated parameters of our credit risk model with that of a conventional default prediction model. We find that accounting for network effects makes a significant contribution to increasing the default prediction power of risk models built specifically for SMEs.

E1397: **Machine learning insights from sparse exchangeable graphs**
Presenter: **Victor Veitch**, Columbia University, United States
Co-authors: Ekansh Sharma, Zacharie Naulet, Daniel Roy
Many tasks in machine learning, e.g. matrix factorization, topic modeling, and feature allocation, can be viewed as learning the parameters of a probability distribution over bipartite graphs. Recent work has introduced the sparse exchangeable (graphex) models as a new family of probability distributions over graphs (in particular, over bipartite graphs). These models offer natural generalizations of many popular approaches to machine learning tasks. Thus the sparse exchangeable models and the associated theory have immediate import for machine learning. We explain some

practical aspects of this connection, with a particular emphasis on the role of subsampling. Further, we introduce sparse exchangeable non-negative matrix factorization as an extended example, which is of interest in its own right.

EO027 Room CLO 203 RECENT APPLICATIONS AND METHODS IN DIRECTIONAL STATISTICS Chair: Eduardo Garcia-Portugues

E0865: **Analyzing infant head flatness and asymmetry using directional surface normal data from a craniofacial 3D model**
Presenter: **Lasse Holmstrom**, University of Oulu, Finland
Co-authors: Ville Vuollo, Henri Aarnivala, Virpi Harila, Tuomo Heikkinen, Pertti Pirttiniemi, Arja Marita Valkama

Human skull deformation is analyzed using the distribution of head normal vector directions computed from a 3D image. Round head shape corresponds to a uniform distribution of normal vector directions and for a symmetrical head, the distribution exhibits symmetry across the plane that divides the head into left and right halves. Severity of head flatness and asymmetry can then be quantified by suitable functionals of the kernel estimate of the normal vector direction density. The density estimates can be visualized using a 2D contour plot of an area-preserving projection of the estimate to a planar disk. The numerical integration needed in evaluating the density functionals can be performed efficiently by employing an evenly spaced Fibonacci grid on the unit sphere. Using image data from 99 infants and clinical deformation ratings made by experts, our approach is compared with some recently suggested methods. The results show that the proposed method performs competitively. We also describe a novel scale space approach to assess the statistical significance of features in a general spherical density function.

E0791: **A generative angular model of protein structure evolution**
Presenter: **Michael Golden**, University of Oxford, United Kingdom
Co-authors: Eduardo Garcia-Portugues, Michael Sorensen, Kanti Mardia, Thomas Hamelryck, Jotun Hein

Recently described stochastic models of protein evolution have demonstrated that the inclusion of structural information in addition to amino acid sequences leads to a more reliable estimation of evolutionary parameters. We present a generative, evolutionary model of protein structure and sequence that is valid on a local length scale. The model concerns the local dependencies between sequence and structure evolution in a pair of homologous proteins. The evolutionary trajectory between the two structures in the protein pair is treated as a random walk in dihedral angle space, which is modelled using a novel angular diffusion process on the two-dimensional torus. Coupling sequence and structure evolution in our model allows for modelling both 'smooth' conformational changes and 'catastrophic' conformational jumps, conditioned on the amino acid changes. The model has interpretable parameters and is comparatively more realistic than previous stochastic models, providing new insights into the relationship between sequence and structure evolution. For example, using the trained model we were able to identify an apparent sequence-structure evolutionary motif present in a large number of homologous protein pairs. The generative nature of our model enables us to evaluate its validity and its ability to simulate and infer aspects of protein evolution conditioned on an amino acid sequence, a related amino acid sequence, a related structure or any combination thereof.

E0847: **A Bayesian network model for linear-circular data**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Ignacio Leguey, Concha Bielza, Pedro Larranaga

In numerous academic fields, it is common that multivariate data include circular observations. Because of the periodic nature of circular observations, a direct application of ordinary Bayesian network techniques could lead to an erroneous result in analysis. We propose a tree-structured Bayesian network model for linear-circular data, namely, data comprising of multiple linear and circular observations. The proposed model is defined using marginals and conditionals of the following three bivariate models: the bivariate normal distribution, a bivariate wrapped Cauchy distribution, and an existing bivariate distribution on the cylinder. There is an efficient algorithm for random generation of the presented Bayesian network model. The mutual information for the joint distributions of parent and child variables can be expressed in a simple and closed form. Maximum likelihood estimation of the parameters is efficient for the marginals and conditionals related to parent and child variables. There are closed-form expressions for the method of moments estimators of the parameters. The presented Bayesian network model can be estimated based on the mutual information via Chow Liu algorithm. A simulation study and an application of the proposed model are given.

E0864: **On some tests for rotational symmetry**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium
Co-authors: Davy Paindaveine, Eduardo Garcia-Portugues

Most commonly-used distributions on unit hyperspheres assume that the data are rotationally symmetric about some direction. The problem of testing for rotational symmetry has been relatively well studied in the circular case. Surprisingly, it has been quite less investigated in the spherical or hyperspherical one. We propose new tests of rotational symmetry on hyperspheres. We show that our tests enjoy many desirable properties. In particular they are locally and asymptotically optimal against new classes of distributions. Our results are illustrated via Monte-Carlo simulations and on a real data related to the study of protein structures.

EO621 Room CLO 204 MIXED LINEAR MODELS WITH APPLICATIONS TO SMALL AREA ESTIMATION Chair: Dietrich von Rosen

E0333: **Poverty mapping in small areas under a two-fold nested error regression model**
Presenter: **Domingo Morales**, University Miguel Hernandez of Elche, Spain
Co-authors: Isabel Molina, Yolanda Marhuenda, Jon Rao

Small area estimation procedures typically provide more reliable poverty estimates than direct methods. These models include area effects to account for the unexplained between-area heterogeneity. When poverty figures are sought at two different aggregation levels, domains and subdomains, it is reasonable to assume a two-fold nested error model including random effects explaining the heterogeneity at the two levels of aggregation. The empirical best (EB) method is introduced for estimating of additive parameters in small areas, under a two-fold model. Under this model, analytical expressions for the EB estimators of poverty incidences and gaps in domains or subdomains are given. The obtained EB estimates of the totals for all the subdomains in a given domain add up to the EB estimate of the domain total. We develop a bootstrap estimator of the mean squared error (MSE) of EB estimators and study the effect on the MSE of a misspecification of the area effects. In simulations, we compare the estimators obtained under the two-fold model with those obtained under models with only domain effects or only subdomain effects, when all subdomains are sampled or when there are unsampled subdomains. The methodology is applied to poverty mapping in counties of the Spanish region of Valencia by gender. Results show great variation in the poverty incidence and gap across the counties from this region, with more counties affected by extreme poverty when restricting ourselves to women.

E0812: **On estimation and prediction in multivariate mixed linear models**
Presenter: **Tatjana von Rosen**, Stockholm University, Sweden

Mixed linear models have been extensively studied due to their intensive use in practical data analysis when the response is univariate. Much attention has been paid on the equality of fixed parameter estimates and the equality of best linear unbiased predictors of random effects under two mixed models with different covariance matrices. The latter can be an important question, for example, in small-area estimation. Nevertheless, these issues have not been considered in the context of multivariate mixed linear models. Therefore, the goal is to extend the results on the equality

of best linear unbiased estimators and best linear unbiased predictors to the multivariate situation. For simlicity, the balanced multivariate mixed linear models will be studied.

E1557: **Testing of multivariate spline growth model**
Presenter: **Tapio Nummi**, University of Tampere, Finland
A new method is presented for testing multivariate growth curves which is based on spline approximation and on F-test. We show how the basic spline regression model can easily be extended to the multiple response case. The method is illustrated using a real data set of multivariate modeling in behavioral cardiology.

E1319: **Small area estimation via a multivariate linear model for repeated measures data**
Presenter: **Dietrich von Rosen**, Swedish University of Agricultural Sciences, Sweden
Co-authors: Martin Singull, Joseph Nzabanita, Innocent Ngaruye
Small area estimation under a multivariate linear model for repeated measures data is considered. The proposed model aims to get a model which borrows strength both across small areas and over time. The model accounts for repeated surveys, grouped response units and random effects variations. Estimation and interpretation of model parameters are discussed. Moreover, prediction of random effects, small area means across time points and per group units are derived.

EO096 Room SH349 MEASURE TRANSPORTATION AND STATISTICS Chair: Eustasio del Barrio

E0235: **Transport based kernel for GP models**
Presenter: **Jean-Michel Loubes**, University of Toulouse, France
Monge-Kantorovich distances, otherwise known as Wasserstein distances, have received a growing attention in statistics and machine learning as a powerful discrepancy measure for probability distributions. We focus on forecasting a Gaussian process indexed by probability distributions. For this, we provide a family of positive definite kernels built using transportation based distances. We provide a probabilistic understanding of these kernels and characterize the corresponding stochastic processes. Then we consider the asymptotic properties of the forecast process.

E0619: **Robust clustering tools based on optimal transportation**
Presenter: **Eustasio del Barrio**, Universidad de Valladolid, Spain
A robust clustering method for probabilities in Wasserstein space is introduced. This new 'trimmed k-barycenters' approach relies on recent results on barycenters in Wasserstein space that allow intensive computation, as required by clustering algorithms. The possibility of trimming the most discrepant distributions results in a gain in stability and robustness, highly convenient in this setting. As a remarkable application we consider a parallelized estimation setup in which each of m units processes a portion of the data, producing an estimate of k-features, encoded as k probabilities. We prove that the trimmed k-barycenter of the $m \times k$ estimates produces a consistent aggregation. We illustrate the methodology with simulated and real data examples.

E0627: **Tools for clustering based on k-barycenters in the Wasserstein space**
Presenter: **Hristo Valdes**, University of Valladolid, Spain
Co-authors: Carlos Bea, Eustasio del Barrio
Cluster analysis addresses the detection of data grouping in data sets. Within this, too vague, description, model-based clustering aims to find particularly shaped groupings -clusters- according to specified distributions. In this setting, the clusters provided by the method are described by probability (often Gaussian) distributions, that can be considered as elements of an abstract space. Particular interest has been deserved by the L2 Wasserstein distance, leading to a rich set-up for developing statistical concepts in a parallel way to those known on Euclidean spaces. This is the case of the k-barycenters, the abstract version of k-means, by large the widest used method in clustering problems, recently introduced in the Wasserstein space even in a robust version. We focus on the application of the (trimmed) Wasserstein k-barycenters to some of the fundamental problems present in cluster analysis. This includes parallelization or stabilization of procedures and even improvement of initial solutions for the algorithms involved in the methods, but we will also pay special attention to the meta-analysis tools arising from this robust aggregation procedure: Stability (or coherence) criteria, applied to the provided aggregation, will give descriptive signs on the number of clusters or on the adequacy of the clustering procedure. We present illustrative examples of the previously mentioned concepts.

E1376: **k-means and optimal transport**
Presenter: **Thibaut Le Gouic**, Institut Mathematiques de Marseille, France
Co-authors: Quentin Paris
In order to apply it to empirical measures, k-means inequalities are studied with view points of optimal transport tools.

CO294 Room MAL 414 MULTIVARIATE VOLATILITY MODELS Chair: Jean-Michel Zakoian

C0395: **On the economic determinants of optimal stock-bond portfolios: International evidence**
Presenter: **Christian Conrad**, Heidelberg University, Germany
Using a modified DCC-MIDAS specification that allows the long-term correlation component to be a function of multiple explanatory variables, the stock-bond correlation in the US, the UK, Germany, France, and Italy is shown to be mainly driven by inflation and interest rate expectations as well as a flight-to-safety during times of stress in financial markets. Based on the new DCC-MIDAS model, we construct stock-bond hedge portfolios and show that these portfolios outperform various benchmark portfolios in terms of portfolio risk. While optimal daily weights minimize portfolio risk, we find that portfolio turnover and trading costs can be substantially reduced when switching to optimal monthly weights.

C1308: **Estimation risk for the VaR of portfolios driven by semi-parametric multivariate models**
Presenter: **Christian Francq**, CREST and University Lille III, France
Co-authors: Jean-Michel Zakoian
Joint estimation of market and estimation risks in portfolios is investigated, when the individual returns follow a semi-parametric multivariatedynamic model and the asset composition is time-varying. Under ellipticity of the conditional distribution, asymptotic theory for the estimation of the conditional Value-at-Risk (VaR) is developed. An alternative method - the Filtered Historical Simulation - which does not rely on ellipticity, is also studied. Asymptotic confidence intervals for the conditional VaR, which allow us to simultaneously quantify the market and estimation risks, are derived. The particular case of minimum variance portfolios is analyzed in more detail. Potential usefulness, feasibility and drawbacks of the two approaches are illustrated via Monte-Carlo experiments and an empirical study based on stock returns.

C1059: **Misspecification tests in conditional covariances for large cross-sectional dimensions**
Presenter: **Bilel Sanhaji**, Paris VIII, France
Co-authors: Thomas Chuffart
Lagrange multiplier tests for nonlinearity in conditional covariances in multivariate GARCH models are proposed. The null hypothesis is the full

BEKK model with variance targeting in which covolatilities of time series are driven by a linear function of their own lags and lagged squared innovations. The alternative hypothesis is an extension of the model in which covolatilities are modeled by a nonlinear function of the lagged squared innovations, represented by an exponential or a logistic transition function. Partial tests are also introduced in order to determine whether the relationship of time series or group of time series is linear or nonlinear. We investigate the size and power of these tests through Monte Carlo experiments, and we provide empirical illustrations in many of which cases these tests encourage the use of nonlinearity in conditional covariances.

C1817: **Consistent pseudo-maximum likelihood estimators and groups of transformations**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Christian Gourieroux, Alain Monfort
In a transformation model $y_t = c[a(x_t, \beta), u_t]$, where the errors u_t are i.i.d, the parameters can be estimated by pseudo-maximum likelihood (PML) method, that is by using a misspecified distribution of the errors, but the PML estimator of β is in general not consistent. We explain how to nest the initial model in an augmented model with more parameters in order to derive consistent PML estimators of appropriate functions of parameter β. The usefulness of the consistency result is illustrated by examples of systems of nonlinear equations, conditionally heteroskedastic models, stochastic volatility, or models with spatial interactions.

CO091 Room MAL 416 NETWORKS AND CAUSALITY Chair: Monica Billio

C0800: **Business shock transmission: A multilayer network perspective**
Presenter: **Monica Billio**, University of Venice, Italy
Co-authors: Roberto Casarin, Luca Rossini
Network structures from a vector autoregressive(VAR) model are extracted to investigate shock transmission effects between variables and cluster the linkages into groups at different layer and at multilayer. A previous multiple shrinkage prior proposed allows us to better understand shock transmission phenomena, the extraction of multilayer and coloured contagion networks and the classification of the linkages strength. The empirical results represent a different point of view on international business cycles providing interesting new findings in the relationship between core and periphery countries and in the centrality of countries within each group and over lags. Finally, the multilayer representation provides evidence of stronger shock transmission contagion effects than in single layers and we show a better identification of the most important and central variables in business shock transmission.

C1155: **Sparse causality networks through regularised regressions**
Presenter: **Michele Costola**, SAFE, Goethe University Frankfurt, Germany
Co-authors: Mauro Bernardi
The aim is to propose a Bayesian approach to the problem of variable selection and shrinkage in high dimensional causal sparse regression models where the regularisation method is an extension of a previous LASSO in a Bayesian framework. The model allows us to extend the pairwise Granger (and quantile) causality in the network estimation by including a large number of institutions which improves the identification of the relationship and maintains at the same time the flexibility of the univariate framework. Furthermore, we obtain a weighted directed network since the adjacency matrix is built "row by row" using for each institutions the posterior inclusion probabilities of the other institutions in the network.

C0785: **Complexity and disagreement in financial networks: A digraph Laplacian approach**
Presenter: **Lorenzo Frattarolo**, Universita Ca Foscari, Italy
Co-authors: Monica Billio, Roberto Casarin, Michele Costola
The aim is to develop new measures of network connectivity, that are Von Neumann entropies, related to complexity of the network, and disagreement persistence index, related to learning and opinion dynamics on the network. Both measures are based on the spectrum of Laplacians defined on directed graphs. We show that the new measures account for global connectivity patterns given by paths and walks of the network. We apply the new measures to a sequence of inferred pairwise-Granger networks. In the application, we employ the proposed measures for the system immunization and early warning for banking crises.

C0875: **The impact of network connectivity on factor exposures, asset pricing and portfolio diversification**
Presenter: **Roberto Panzica**, Goethe University House of finance, Italy
Co-authors: Loriana Pelizzon, Massimiliano Caporin, Monica Billio
The classic factor based asset pricing model is extended by including networklinkages in linear factor models. We assume that the network linkages are exogenously provided. This extension allows a better understanding of the causes of systematic risk and shows that (i) network exposures act as an inflating factor for systematic exposure to common factors and (ii) the diversification power is reduced by the presence of network connections. Moreover, we show that in the presence of network links a misspecified traditional linear factor model presents residuals that are correlated and heteroskedastic. We support our claims with an extensive simulation experiment

CO071 Room MAL 421 TOPICS IN NON-STATIONARY TIME SERIES Chair: Martin Wagner

C1298: **Residual based detection of market dislocations**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner
Multivariate monitoring procedures are developed to detect market dislocations, e.g., in covered or uncovered interest parities. The simultaneous consideration of multiple cointegration relationships, in the above example corresponding to different maturities, is intended to increase the monitoring performance compared to univariate approaches. The procedures are based on monitoring variants of well-known non-parametric cointegration rank statistics. In case the parameters of the considered relationship have to be estimated rather than assumed to be known from theoretical arguments, modified least squares estimation techniques need to be used for asymptotically nuisance parameter free limiting distributions of the monitoring procedures. The different monitoring procedures are compared by means of simulations and applied to interest rate parity data for eight countries.

C1578: **The asymptotic validity of "standard" FM-OLS estimation and inference in cointegrating polynomial regressions**
Presenter: **Martin Wagner**, Technical University Dortmund, Germany
Co-authors: Oliver Stypka, Peter Grabarczyk, Rafael Kawka
Estimation and inference in cointegrating polynomial regressions is considered which includes regressions with deterministic variables, integrated processes and their powers as explanatory variables. The stationary errors are allowed to be serially correlated and the regressors are allowed to be endogenous. The main result shows that estimating such relationships using the fully modified OLS approach developed for linear cointegrating relationships by incorrectly considering all integrated regressors and their powers as integrated regressors leads to the same limiting distribution as the fully modified type estimator developed for cointegrating polynomial regressions. A key ingredient for the main result are novel limit results for kernel weighted sums of properly scaled nonstationary processes involving scaled powers of integrated processes. Even though the simulation

results indicate performance advantages of the latter estimator that are partly present even in large samples, the results drastically enlarge the usability of the former estimator as implemented in many software packages.

C1620: **Testing for a stochastic unit root-type process using Chebyshev time polynomials approximation**
Presenter: **Julio Angel Afonso-Rodriguez**, University of la Laguna, Spain
Many macroeconomic and financial time series seem to be well characterized by some periods of stationary or nonstationary behaviour, combined with some other periods of explosiveness or exuberance which resembles the bubble phenomenon. Among the different existing possibilities to account for this variety of behaviours, we focus on the case of a stochastic unit root (STUR) type of processes, which includes the pure STUR, the bilinear unit root (BLUR) and the threshold autoregressive STUR (TARSUR) processes. Our main interest is on testing for the existence of this type of processes as the alternative hypothesis to the null of a fixed unit root process. To that end we propose the novelty approach based on introducing a number of trigonometric terms, given by Chebyshev time-polynomials, in the auxiliary regression of the testing procedure to capture the time-varying property of the autoregressive unit root component. We obtain the limiting null and the local-to-the alternative distribution of a generalized pseudo-F ratio test statistic, which have several appealing properties over existing test statistics particularly under serially correlated error terms, and show that it has good power behaviour against any of the particular forms of the STUR-type alternative.

C1581: **GLS estimation and confidence sets for the date of a single break in models with trends**
Presenter: **Yicong Lin**, Maastricht University, Netherlands
Co-authors: Eric Beutner, Stephan Smeekes
The aim is to present the asymptotic results on generalized least squares (GLS) estimation of time series models with a single break in level and/or trend at some unknown date where the errors are assumed to be stationary. As GLS estimators, relying on the unknown inverse covariance matrix of the errors, are usually infeasible, we estimate the unknown inverse matrix after doing the modified Cholesky decomposition. Then, we construct feasible GLS estimators. Based on inverting a sequence of likelihood-ratio tests, we construct valid confidence sets for the timing of the break and compare these with existing methods. Extensive Monte Carlo simulations demonstrate that our proposed method has good finite sample properties in terms of either the coverage rates or the lengths of the confidence sets.

CO310 Room MAL 532 TOPICS IN NONPARAMETRICS Chair: Artem Prokhorov

C0566: **Smooth minimum distance inference for semiparametric partially linear regressions with Box-Cox transformation**
Presenter: **Daniel Becker**, University of Bonn, Germany
Co-authors: Alois Kneip, Valentin Patilea
A semiparametric partially linear model with Box-Cox transformed dependent variable is studied. Transformation regression models are widely used in applied econometrics to avoid misspecification. In addition, a partially linear semiparametric model is an intermediate strategy that tries to balance advantages and disadvantages of a fully parametric model and nonparametric models. A combination of transformation and partially linear semiparametric model is, thus, a natural strategy. The model parameters are estimated by the so called smooth minimum distance (SmoothMD) estimator. SmoothMD is suitable for models defined by conditional moment conditions and allows the variance of the error terms to depend on the covariates. The asymptotic behavior of the new SmoothMD estimator is studied under general conditions and new inference methods are proposed. A simulation experiment illustrates the performance of the methods for finite samples. An extension of the model to models with endogenous variables is considered as well.

C0896: **Inference with Hamiltonian sequential Monte Carlo**
Presenter: **Martin Burda**, University of Toronto, Canada
Co-authors: Remi Daviet
A key problem with traditional MCMC methods using Metropolis-Hastings or Gibbs sampling is slow mixing of the chain under multimodality, concentrated mass or a complex shape of the objective function. A new Monte-Carlo method is proposed by combining the advantages of Sequential Monte Carlo simulators and Hamiltonian Monte Carlo simulators. The result is robust to multimodality and complex shapes of the objective function. We further enhance the method with a kernel-based resampling step to enhance robustness and efficiency. We present several challenging simulated examples.

C0805: **Two-sample estimation as an alternative to instrumental variable estimation in the presence of omitted variables**
Presenter: **Masayuki Hirukawa**, Setsunan University, Japan
Co-authors: Irina Murtazashvili, Artem Prokhorov
When conducting regression analysis, econometricians often face the situation in which some regressors are unavailable in the data set at hand (e.g., an ability measure in wage regression). They typically treat the missing regressors as omitted, and for consistent estimation of model parameters, they attempt to find valid instruments for the regressors that are suspected to be endogenous due to their possible correlation with the omitted regressors. We argue that we can also estimate parameters by combining the original data set with another one containing the 'missing' regressors. A consistent, two-sample estimator is proposed as an alternative to the matched sample indirect inference estimator, and its asymptotic properties are explored. Moreover, we refer to dimension reduction methods of matching variables and conduct Monte Carlo simulations.

C0990: **Efficient semiparametric copula estimation of regression models with endogeneity**
Presenter: **Kien C Tran**, University of Lethbridge, Canada
Co-authors: Artem Prokhorov, Mike Tsionas
An efficient one-step semiparametric copula estimation approach is proposed for estimating regression models with endogenous regressors. In particular, the proposed approach is based on sieve maximum likelihood procedure. We derive the asymptotic properties of the proposed estimator, and show that the proposed approach is efficient in the sense that it achieves semiparametric efficiency bound. Monte Carlo simulations show that the proposed approach perform well in finite sample.

CO518 Room MAL 538 ROBUST PORTFOLIO ANALYSIS Chair: Winfried Pohlmeier

C0224: **Deciding with judgment**
Presenter: **Simone Manganelli**, European Central Bank, Germany
Non sample information is hidden in frequentist statistics in the choice of the hypothesis to be tested and of the confidence level. Explicit treatment of these elements provides the connection between Bayesian and frequentist statistics. A frequentist decision maker starts from a judgmental decision (the decision taken in the absence of data) and moves to the closest boundary of the confidence interval of the first order conditions, for a given loss function. This statistical decision rule does not perform worse than the judgmental decision with a probability equal to the confidence level. For any given prior, there is a mapping from the sample realization to the confidence level which makes Bayesian and frequentist decision rules equivalent. An application to a portfolio choice between cash and the EuroStoxx50 index illustrates the practical advantages of frequentist decision rules.

C0219: **Portfolio pretest estimation with machine learning**
Presenter: **Ekaterina Kazak**, University of Konstanz, Germany
Co-authors: Winfried Pohlmeier
The general idea of the pretest estimation is to choose optimally between various competing estimators based on a testing outcome. However, the optimal decision rule may not rely on conventional choices for the significance level. Under certain circumstances it might be reasonable to select a lower significance level (higher probability of a Type I error) than a conventional one in order to increase the probability of rejecting the benchmark strategy. This problem arises in portfolio analysis, when the investor has to decide between two or more alternative portfolio strategies in the presence of low powered performance tests when truly superior strategies are rejected in favor of the benchmark strategy. The optimal size for the pretesting strategy can often not be derived and may change over time in time-series settings. We develop a data driven approach for choosing an optimal significance level for pretest estimators. We show that the bagged pretest estimator performs exceptionally well, especially when combined with adaptive smoothing. The resulting strategy allows for a flexible and smooth switch between the underlying strategies and is shown to outperform the corresponding stand-alone strategies. Our learning pretest estimation technique is shown be a strong competitor to alternative regularization based portfolio estimators.

C0603: **A portfolio perspective on the multitude of firm characteristics**
Presenter: **Alberto Martin Utrera**, Lancaster University Management School, United Kingdom
The aim is to investigate which characteristics matter jointly for an investor who cares not only about average returns but also about portfolio risk, transaction costs, and out of sample performance. We find only a small number of characteristics (six) are significant without transaction costs. With transaction costs, the number of significant characteristics increases to 15 because the trades in the underlying stocks required to rebalance different characteristics often net out. We show investors can identify combinations of characteristics with abnormal out-of-sample returns net of transaction costs that are not fully explained previous factors.

C1619: **A new non-Gaussian factor GARCH model**
Presenter: **Patrick Walker**, University of Zurich, Switzerland
Co-authors: Marc Paolella, Pawel Polak
A new orthogonal factor GARCH model for a multivariate set of skewed, heavy-tailed asset returns is proposed. Conditional returns are modeled by a multivariate generalized hyperbolic (MGHyp) distribution and the covariance matrix specification makes use of the eigendecomposition of principle component analysis (PCA). Due to the mixing structure of the MGHyp distribution, the filtered residuals are Gaussian and common market shocks to asset volatilities, which manifest themselves in leptokurtic tails, are accounted for. The leading eigenvalues of the orthogonal decomposition, representing those unobserved statistical factors that explain most of the assets' covariances, are endowed with univariate GARCH dynamics, while the remaining eigenvalues are assumed constant over time in order to preserve invertibility. Joint maximum likelihood estimation of all model parameters is carried out by a fast expectation conditional maximization either (ECME) algorithm. An application to portfolio optimization with daily rebalancing shows the superiority of the new model compared to several benchmark models and the naive diversification strategy. In addition, we observe lower portfolio turnover of our new model compared to multivariate GARCH models with constant conditional correlations (CCC), translating into possibly lower transaction costs.

CO348 Room MAL 539 ECONOMETRICS AND ITS APPLICATION Chair: Yong Song

C0244: **Information disclosure discounts in takeover auctions**
Presenter: **Dong-Hyuk Kim**, University of Queensland, Australia
Motivated by the concerns on the confidential information disclosure and the practice of bidder regulation in takeover auctions, we develop a structural auction model that approximates the takeover procedure to study the effect of confidential information disclosure on bidders' behavior and seller's revenue. We first derive a bidders' optimal bidding strategy and provide a condition for the existence of a symmetric monotone BNE. Then, we discuss the identification of the model with and without unobserved target heterogeneity. Finally, we estimate the model with the data on takeover auctions of U.S. public companies.

C0255: **Real economic growth and oil price shocks: A Bayesian approach**
Presenter: **Yong Song**, University of Melbourne, Australia
Co-authors: John Maheu, Qiao Yang
The predictability of U.S. economic growth by oil price is revisited. Existing works focus on point forecasts through the lens of linear models. We apply nonlinear models to investigate the impact of oil price on higher moments of the growth. In addition to point forecasts, we evaluate density forecasts as a measure of the overall predictive power.

C0318: **High dimensional M-estimation for large panel data models**
Presenter: **Huanjun Zhu**, Xiamen University, China
M-estimation for large panel data model with high dimensional parameters is studied. We investigate the influence of various of dimensions on the statistical inference for M-estimation, including cross-sectional dimension, time series length and parameter dimension. The rate of convergence and asymptotic distribution of the provided M-estimator are established. Moreover, for easy practice, a good estimator for the asymptotic variance in the asymptotic distribution is constructed and its consistency is discussed in theoretical form. All kinds of simulations illustrate the validity of M-estimation for heavy-tailed data and data with outliers. Empirical application on stock returns is also provided.

C1410: **Pooled infinite hidden Markov models and the Taylor rule: An empirical study on the U.S. monetary policy**
Presenter: **Zhuo Li**, University of Melbourne, Australia
Co-authors: Yong Song
A Bayesian non-parametric technique, the Infinite Hidden Markov Model (IHMM), is applied in estimating real-time Taylor Rule from 1969q1 to 2008q4. In contrast with the conventional k states Markov Switching model, where k has to be preassigned, IHMM estimates the k within the entire natural number domain, which explores the regime uncertainty in a rich content. Rather than embedding the Federal Fund rate in a single uni-variate equation, we parallel estimate the policy function with 6 alternative forms, each of them stands for a specific policy horizon that ranges from backward- to forward-looking ones. In order to capture the dynamic variations among policy horizon, we further use a Dynamic pooling technique among the underlying policy rules. Merging the predictive likelihood based pooling method with IHMM, we are able to analysis the policy horizon as well as the regime uncertainties in a parsimonious way without stringent assumptions. In our findings, for disjoint time intervals, we draw both comparable and different conclusions with respect to the existing literature. Moreover, taking the marginal likelihoods as an evidence, the results suggest that Dynamic pooling is superior than the Static counterpart or the standard Bayesian Model Averaging in tackling policy horizon uncertainty.

CO254 Room MAL 540 QUANTITATIVE INVESTING Chair: Serge Darolles

C0382: **The smart vega factor-based investing: Disentangling risk premia from implied volatility smirk**
Presenter: **Anmar Al Wakil**, University Paris-Dauphine, PSL Research University, France
The way is paved for new option-based volatility strategies genuinely built on factor-based investing. Since market option prices reflect uncertainty, the pricing discrepancy between the physical and the risk-neutral distributions, i.e. the fair price of moments, is exploited. From an economic perspective, the level, slope, and convexity associated to the implied volatility smirk quantify the departure of the returns probability distribution from the lognormal distribution. Subsequently, the so-called "Smart Vega investing" proposes option-based replication strategies mimicking the volatility, skewness, and kurtosis risk premia in the form of divergence swap contracts, tradeable at moderate transaction costs in incomplete option markets. Extending a quadratic approximation, an explicit representation of the implied volatility smirk function, conveniently expressed as a combination of tradeable time-varying risk premia that reward for bearing higher-order risks, is exploited. Furthermore, these theoretical underpinnings are tested on S&P 500 and VIX options, under strongly skewed leptokurtic distributions.

C1379: **Managing hedge fund liquidity risks**
Presenter: **Serge Darolles**, Paris Dauphine, France
Co-authors: Guillaume Roussellet
Hedge fund optimal portfolios are studied in the presence of market and funding liquidity risks. We consider a two-period economy with a single hedge fund. The fund has access to cash, which is available every period, and to an illiquid asset, which pays off only at the end of the second period. Funding liquidity risk takes the form of a random proportion of the fund's assets under management being withdrawn by clients in period one. The fund can then liquidate a part of the illiquid position by bidding on a secondary market where a random haircut on the effective selling price is applied. We solve the allocation problem of the fund and find its optimal portfolio. Whereas the cash buffer is monotonously decreasing in the secondary market liquidity, we show that the fund's default probability is bell-shaped.

C1021: **Risk-based allocation for illiquid and alternative investments**
Presenter: **Emmanuel Jurczenko**, Ecole Hoteliere de Lausanne, Switzerland
Co-authors: Jerome Teiletche
A generalized risk-based investing framework is proposed, which makes it possible to deal in a simple and flexible way with various risks beyond volatility, namely valuation, asymmetry, tail and illiquidity risks. We empirically illustrate the methodology by proposing a risk-based strategic allocation for a multi-asset portfolio made of traditional and alternative assets with different degrees of liquidity.

C0961: **Styles of private equity funds**
Presenter: **Elise Gourier**, Queen Mary University of London, United Kingdom
Co-authors: Ludovic Phalippou, William Goetzmann
A methodology is provided to statistically identify the different styles of private equity funds. Styles are defined based on the dynamics of the funds' return processes, which are estimated from observed cash flows and reported net asset values. We analyze the risk-return profile of each style and investigate whether they match the self-declared categories of private equity funds. The results give insights on the diversification benefits of the different types of alternative investments.

CO108 Room MAL 541 COMPUTATIONAL METHODS IN FINANCIAL ECONOMETRICS Chair: Sandra Paterlini

C1042: **Sparse indirect inference**
Presenter: **Paola Stolfi**, Roma Tre University, Italy
Co-authors: Mauro Bernardi, Lea Petrella
Indirect inference methods are simulation-based procedures for estimating the parameters of an intractable model using an alternative auxiliary model which is computationally simpler. Such methods have received a lot of attention in literature, but no attempt has been made to deal with large dimensional problems. We introduce the sparse indirect inference estimator by adding the Smoothly Clipped Absolute Deviation penalty. We establish consistency and asymptotic normality of the proposed estimator and we show that the Sparse-Indirect Inference estimator enjoys the oracle properties under mild regularity conditions. The method is applied to estimate the parameters of large dimensional Non-Gaussian dynamic regression models.

C1062: **Sparse precision matrices for minimum variance portfolios**
Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Sandra Paterlini, Rosella Giacometti
Financial crises are typically characterized by highly positively correlated asset returns due to the simultaneous distress on almost all securities, high volatilities and the presence of extreme returns. In the aftermath of the 2008 crisis, investors were prompted even further to look for portfolios that minimize risk and can better deal with estimation error in the inputs of the asset allocation models. The minimum variance portfolio a la Markowitz is considered the reference model for risk minimization, due to its simplicity in the optimization as well as its need for just one input estimate: the inverse of the covariance estimate, or the so-called precision matrix. We propose a data-driven portfolio framework that relies on two regularization methods, glasso and tlasso. They provide sparse estimates of the inverse of the covariance matrix by penalizing the 1-norm of the precision matrix relying on asset returns normality or t-Student assumptions, respectively. Simulation and actual data results support the proposed methods compared to state-of-art methods, such as random matrix and Ledoit-Wolf shrinkage.

C0703: **Modeling the U.S. oil market using heterogeneous interacting agents**
Presenter: **Christoph Funk**, Justus-Liebig-University Giessen, Germany
Co-authors: Johannes Lips
The recent decline in the price of oil since June 2014 and the increase in U.S. shale oil production gives an interesting environment for the use of heterogeneous interacting agents. We model the oil producing sector in the U.S. on a firm level by using a production function based on the endowment of capital assets (oil wells and fields) and a fixed labor input. Thereby, we are interested in investigating the interaction between shale oil and conventional oil producers and thus, the consequences for the real price of oil. We expect two different outcomes based on the initial endowment of capital assets. As the overall production level increases and prices decline, shale oil producers might not lower their production in order to satisfy capital investor's needs. Hence, shale oil producers show a trend toward over-leveraging, a high risk of insolvency and a tendency of excess production. In addition, large capital companies are less dependent on oil price fluctuations and are more likely to survive a long lasting decline in the price of oil. The model calibration will be evaluated using firm data for the U.S. oil market.

C1354: **Creating (parsimonious) banking networks**
Presenter: **Sandra Paterlini**, European Business School Germany, Germany
Co-authors: Ben Craig, Dietmar Maringer
The level of interconnectedness of financial institutions and the network topology are central in determining the resilience of the system and the channels of shock propagation, as well as understanding how individual institutions take decisions and form links. Still, data are typically scarce and when available, only at an aggregate level. We introduce a simple yet effective decreasing marginal cost optimization model to estimate the interbank market network. By relying on an improved optimization algorithm from transport theory, we show that our model not only allows us to capture some stylized facts of the interbank networks, such as the core-periphery structure, but it also provides insights on the effect of the marginal distributions of assets and liabilities, as well as the consequences of potential heterogeneous preferences among banks.

CC710 Room MAL 151 CONTRIBUTIONS IN MACROECONOMETRICS Chair: Frederic Karame

C0332: **Particle filtering with Dynare**
Presenter: **Frederic Karame**, Le Mans University, France
Co-authors: Stephane Adjemian
Since the 2000s, the literature on dynamic stochastic general equilibrium (DSGE) models has evolved towards the use of higher order local approximations. This shift towards nonlinear reduced form solution has implications for the estimation of DSGE models. One now needs to use nonlinear methods, such as Sequential Monte Carlo (also known as particle filtering) to compute the likelihood of the approximated model. The aim is to present some of these developments and their implementation in Dynare, the free, user-friendly and intuitive software platform that handles a wide class of economic models, in particular DSGE and OLG models.

C1370: **Monetary policy regime shifts and inflation persistence in the UK**
Presenter: **Shayan Zakipour-Saber**, Queen Mary University, United Kingdom
Two monetary policy regimes in the United Kingdom between 1960-2009 are identified by estimating an open economy general equilibrium model in which parameters of a Taylor-type policy rule and volatilities of structural economic shocks follow independent Markov processes (MS-DSGE). Both regimes actively respond to movements in inflation, by adjusting the nominal interest rate more than one-for-one. One regime is significantly more active and is shown to be in place at the beginning of the 1960s and after 1992, which coincides with the period when the Bank of England explicitly adopted a policy of inflation targeting. Inflation persistence is measured as the fraction of the total variation of inflation from a constant target that is due to past shocks. We find a small but significant decrease in the inflation persistence of monetary policy regimes that respond more actively to inflation.

C1365: **Phillips curves**
Presenter: **Luca Onorante**, European Central Bank, Germany
Co-authors: Shayan Zakipour-Saber, Laura Moretti
A robust estimation of the Phillips curve is performed, taking into account the different specifications existing in the literature, the uncertainty in the measurement of variables and the potential non-linearities and structural changes deriving from the Great Recession. We extend the existing literature by adding model specifications, accounting for different measures of inflation expectations and external factors, and choosing different specifications on the basis of their out-of-sample forecasting performance. Additionally, we attempt at identifying the most important determinants of inflation over the sample. Finally, we explore the consequences of our findings for monetary policy: we overall conclude that there is scope for action of monetary policy, but the objective of inflation close but below 2 percent will also crucially depend on the (largely exogenous) developments of food and energy prices, and most crucially on expectations.

C1528: **Sovereign credit rating mismatches**
Presenter: **Antonio Afonso**, ISEG-UL - University of Lisbon, Portugal
The factors behind split ratings in sovereign credit ratings from different agencies are studied for the period 1980-2015. We employ random effects ordered and simple probit approaches to assess the explanatory power of different macroeconomic, government and financial variables. Our results show that structural balances and the existence of a default in the last ten years were the least significant variables whereas the level of net debt, budget balances, GDP per capita and the existence of a default in the last five years were found to be the most relevant variables explaining rating mismatches across agencies. For speculative-grade ratings, we also find that a default in the last two or five years decreases the rating difference between S&P and Fitch. For the positive rating difference between S&P and Moodys for investment-grade ratings, an increase in external debt leads to a smaller rating gap between the two agencies.

CC708 Room MAL 152 CONTRIBUTIONS IN BAYESIAN METHODS Chair: Jim Griffin

C1531: **On the effects of the monetary policy on the income inequality in Japan: Evidence from grouped data**
Presenter: **Kazuhiko Kakamu**, Kobe University, Japan
Co-authors: Martin Feldkircher
The effects of the monetary policy on the income inequality in Japan is examined by using a novel econometric approach that jointly estimates the Gini coefficient based on micro-level grouped data of households and the dynamics of macroeconomic quantities. Our results indicate different effects on income inequality depending on the monetary policy measure under consideration: A traditional rate increase decreases income inequality, whereas a reduction of asset purchases leads to more inequality. Movements of inflation expectations and equity prices might account for theses differences.

C1808: **Nested error regression model with non ignorable missing values**
Presenter: **Hiromasa Tamae**, University of Tokyo, Japan
Co-authors: Shonosuke Sugasawa, Tatsuya Kubokawa
In the context of small area problem, small sample size in each segment, nested error regression model is a powerful tool to build a stable predictor. This, however, assumes that the samples are drawn randomly; seldom achieved in real data survey. Caring selection bias, small area problem can be seen by another aspect: severe missing problem in each area. Unless the missing mechanism is completely random, conventional model does not work well for an accurate inference. We propose a nested error regression model cooperating with non ignorable missing value in which the missing mechanism explicitly expressed as probit regression relating binary observation variables with continuous response variable. For estimation of the model parameters, a Bayesian inference is suggested by setting non-informative prior distributions on the model parameters. To conduct Markov Chain Monte Carlo methods, we construct the Gibbs sampling method for all the full conditional distributions. The proposed model is compared with the standard nested error regression model through numerical simulations.

C1671: **Exchange rate predictive densities: An application of stochastic model specification search**
Presenter: **Emi Mise**, University of Leicester, United Kingdom
Co-authors: Anthony Garratt, Emi Mise

The predictive densities for exchange rates computed from TVP-VECMs incorporating stochastic volatility (SV) are evaluated. The long-run components of the VECMs consider well-known sets of fundamentals-based models of exchange rates, including purchasing power parity, uncovered interest parity, monetary fundamentals as well as Taylor rules. We compute out-of-sample predictive densities by Bayesian model averaging using the posterior model likelihood as model weights. In order to address the well-documented problem of over-parameterisation leading to poor predictive performance, a prior shrinkage procedure is adopted. Various methods of prior shrinkage have been proposed in the literature. We proceed by comparing the predictive densities of TVP-VECM with SV, with and without the application of the shrinkage method. This approach permits a significantly more parsimonious model in which some of the parameters are restricted not to vary over time. In our application, monthly data for U.K. Sterling against U.S. dollar, over the 40-year period after the Bretton-Woods system is used. Our forecast evaluation uses both statistical and economic criteria.

C1449: **Factor state-space models for high-dimensional realized covariance matrices of asset returns**
Presenter: **Bastian Gribisch**, University of Cologne, Germany
Co-authors: Jan Patrick Hartkopf, Roman Liesenfeld

A flexible matrix-variate state-space model is proposed for the prediction of high-dimensional covariance matrices of asset returns based on observable common factors such as the three Fama-French factors. In this model the latent joint covariance matrix of the assets and the common factors is observed through the corresponding realized covariance matrix with a measurement density which is assumed to be a central Wishart. By imposing a factor structure the covariance matrix of the assets is decomposed into a low-rank component driven by the covariance matrix of the factors and a diagonal residual covariance matrix. In this decomposition we allow for dynamic variation in the factor and residual covariance as well as in the factor loadings. This decomposition translates into a convenient factorization of the Wishart measurement density which greatly simplifies statistical inference in high-dimensional applications. The proposed model can be analyzed using Bayesian MCMC procedures and exploiting computational parallelization techniques enables a fast, scalable and numerically precise statistical inference. An empirical application to realized covariance matrices for 60 NYSE traded stocks shows that the model performs very well in- and out of sample.

CG006 Room MAL 402 CONTRIBUTIONS IN NOWCASTING AND BIG DATA Chair: Juan-Pablo Ortega

C1551: **A real time investigation of short-term forecasting**
Presenter: **Alain Galli**, Swiss National Bank, Switzerland
Co-authors: Rolf Scheufele, Christian Hepenstrick

Nowcasting methods to assess the current state of the economy are often evaluated in terms of their out-of-sample forecast performance for quarterly GDP. However, such an assessment heavily depends on the point in time, i.e. the information set, for which such an evaluation is conducted. Furthermore, it ignores that the target variable - the first release of quarterly GDP - is highly prone to revisions. Therefore, we assess such methods in an alternative way and focus on the convergence of each models forecast week by week to its current quarter nowcast at the time when the first estimate of GDP is published. In other words, we answer the question of what can be known when. Furthermore, we show that, for the case of Switzerland, within a period of two years after the respective quarter has ended, the real-time assessment of economic developments based on nowcasting approaches outperform first and subsequent GDP releases in terms of prediction ability for final GDP.

C1596: **Nowcasting firm's financial standing using global inter-firm relationships**
Presenter: **Takayuki Mizuno**, National Institute of Informatics, Japan
Co-authors: Kenta Yamada

The main factor in a companys fundamentals is financial variables. However, it is not always easy to obtain information about them. The financial variables are discretely published quarterly as settled-account reports, whereas stock prices change at speeds of one hundredth of a second in the stock market. To measure the gap between asset price and fundamentals, we must always nowcast company performance in the time between one settlement report and the next. We establish an estimation method of financial variables with high estimation accuracy by using global interfirm relationships. There are many companies with similar businesses, sales areas and sizes. Such companies have business connections. Because the settlement dates are different for many companies, we can estimate unreported financial variables of a particular company from ones already reported by similar companies. We investigate whether this can be achieved using a unique dataset that covers a list of business connections and a highly structured companys revenues by geographic and business segments. Through the analysis of random forests, we found that common shocks affecting one company can be measured from revenues of companies that operate in the same area.

C1766: **News and consumer card payments**
Presenter: **Libero Monteforte**, Bank of Italy, Italy
Co-authors: Guerino Ardizzi, Simone Emiliozzi, Juri Marcucci

Central banks and policy makers are strongly interested in finding ways to measure the impact of policy uncertainty and cyber-security related news on economic activity. We investigate how consumers consumption (saving) decisions and payment instrument choices are influenced by these two important sources of news: the ones concerning economic policy uncertainty and those related to payment card frauds. We use a unique daily data set, for the Italian economy, with payment data recorded electronically through clearing and settlement circuits managed by the Bank of Italy, and coupled with indicators of news derived using big data techniques. We provide a quantitative assessment of how these news shape consumers behaviour and their preferences for liquidity at a daily frequency.

C1433: **Nowcasting private consumption: Traditional indicators, uncertainty measures, and the role of internet search query data**
Presenter: **Alberto Urtasun**, Banco de España, Spain
Co-authors: Javier J Perez

The aim is to nowcast quarterly private consumption. We estimate a suite of mixed-frequency models on a real-time database for the case of the Spanish economy, and conduct out-of-sample forecasting exercises to assess the relevant merits of different groups of indicators. The selection of indicators is guided by the standard practice (hard and soft indicators), but also expand this practice by looking at non-standard variables, namely: (i) a suite of proxy indicators of uncertainty, calculated at the monthly frequency; (ii) two additional sets of variables that are sampled at a much lower frequency: Credit card transactions and indicators based on search query time series provided by Google Trends. The latter set of indicators isbased on factors extracted from consumption-related search categories of the Google Trends application. We also illustrate how Google data (sampled at a frequency higher than monthly) can be instrumental to perform event studies, by looking at possible anticipation effects related to VAT increases.

CG080 Room MAL 415 CONTRIBUTIONS IN VALUE-AT-RISK Chair: Alexandra Dias

C1618: **Multivariate extensions of the ACD peaks-over-threshold method for forecasting value at risk**
Presenter: **Katarzyna Bien-Barkowska**, Warsaw School of Economics, Poland

A new dynamic peaks-over-threshold (POT) model is proposed for extreme events in financial markets. The random times when the sizes of negative financial returns exceed given threshold are modeled in line within the marked point process theory, where the marks correspond to the magnitudes of extreme losses. We develop a multivariate version of the autoregressive conditional duration (ACD) model, where the conditional intensity of extreme negative returns has not only the self-exciting structure, but also the cross-exciting structure, since it can instantaneously react to the time-varying covariates such as large positive returns or volatility peaks. In our approach the observed times of all these intervening events can accelerate or decelerate the awaited occurrence of extreme losses. We apply the extended multivariate version of the ACD model to six major stock indexes and show that it outperforms the standard ACD-based POT methods for forecasting value-at-risk and expected shortfall.

C1629: **Measurement of common risk factors: A panel quantile regression model for returns**
Presenter: **Frantisek Cech**, UTIA AV CR vvi, Czech Republic
Co-authors: Jozef Barunik

The aim is to investigate how to measure common market risk factors using newly proposed Panel Quantile Regression Model for Returns. By exploring the fact that volatility crosses all quantiles of the return distribution and using penalized fixed effects estimator we are able to control for otherwise unobserved heterogeneity among financial assets. Direct benefits of the proposed approach are revealed in the portfolio Value–at–Risk forecasting application, where our modeling strategy performs significantly better than several benchmark models according to both statistical and economic comparison. In particular Panel Quantile Regression Model for Returns consistently outperforms all the competitors in the 5% and 10% quantiles. Sound statistical performance translates directly into economic gains which is demonstrated in the Global Minimum Value–at–Risk Portfolio and Markowitz-like comparison. Overall results of our research are important for correct identification of the sources of systemic risk, and are particularly attractive for high dimensional applications.

C1598: **New indicators in systemic risk analytics: Theory and applications**
Presenter: **Mario Maggi**, University of Pavia, Italy
Co-authors: Silvia Figini, Pierpaolo Uberti

A novel class of indicators to forecast financial crises is presented. The proposal can also be viewed as a supplementary measure in financial systemic risk analysis. The family of indicators introduced is applied to different stock market indexes to assess the predictive ability to detect tensions in the financial market. Moreover, a comparison with alternative measures of systemic risk confirms the predictive ability and shows that our indicator's predicting power outperforms other measures proposed in the literature.

Sunday 17.12.2017 16:35 - 18:15 Parallel Session J – CFE-CMStatistics

EI690 Room CLO B01 CURRENT ISSUES IN CAUSAL INFERENCE FOR MULTIPLE MEDIATORS Chair: Bernard Rachet

E0187: **Explaining the total effect in the presence of multiple mediators and interactions**
Presenter: **Linda Valeri**, Harvard Medical School and McLean Hospital, United States
Mediation analysis allows decomposing a total effect into a direct effect of the exposure on the outcome and an indirect effect operating through a number of possible hypothesized pathways. A recent study has provided formal definitions of direct and indirect effects when multiple mediators are of interested. Parametric and semi-parametric methods to estimate path-specific effects have also been described. Investigating direct and indirect effects with multiple mediators can be challenging in the presence of multiple exposure-mediator and mediator-mediator interactions. Three main contributions are provided: 1) we obtain counterfactual definitions of interaction terms when more than one mediator is present; 2) we derive a decomposition of the total effect that unifies mediation and interaction when multiple mediators are present; and 3) we illustrate the connection between our decomposition and the 4-way decomposition of the total effect introduced in the context of a single mediator. The framework applies to continuous, categorical or time-to-event outcomes. We illustrate the properties of the proposed framework for multiple mediators and interactions, in a secondary analysis of a pragmatic trial for the treatment of schizophrenia. We employ the decomposition to investigate the complex interplay of side-effects and psychiatric symptoms trajectories in explaining the effect of antipsychotics on social functioning in schizophrenia patients.

E0171: **Time-to-event mediation analysis with repeatedly measured mediators**
Presenter: **Stijn Vansteelandt**, Ghent University and London School of Hygiene and Tropical Medicine, Belgium
The LEADER trial found protective cardiovascular effects of liraglutide compared to placebo in patients with Type II diabetes and high cardiovascular risk. Effects were also found on glycated hemoglobin, body weight, blood pressure and heart rate, thereby raising the question to what extent these potential pathways my explain liraglutide's protective effect. We will explain how we addressed this question by expanding modern techniques from causal mediation analysis. In particular, we will show how to identify and infer the natural indirect of liraglutide on the time to major cardiovascular events via the repeatedly measured glycated hemoglobin levels. The considered proposal addresses complications due to patients dying before the mediator is assessed, due to the mediator being repeatedly measured, and due to post-treatment confounding of the effect of glycated hemoglobin by other mediators, which makes mediation analysis a challenging enterprise.

E0172: **Mediation analysis with high-dimensional mediators**
Presenter: **Rhian Daniel**, Cardiff University, United Kingdom
In many modern biomedical applications, interest lies in decomposing the effect of an exposure, eg a genetic variant, on an outcome, eg cardiovascular disease, into its effect via a large number of mediators, eg blood protein and metabolite measures. Such an endeavour poses formidable methodological challenges. First, the mediators are likely to be highly-correlated according to an unknown causal structure, including unmeasured common causes of one mediator and another. Second, the identification of the usual natural path-specific effects in such a setting would rely on many "cross-world independence" assumptions, which are impossible to justify. Third, the usual parametric regression estimation approaches would rely on a huge number of (uncheckable, in practice) modelling assumptions. We propose that the first two problems be overcome by focusing on interventional multiple mediator effects and the third by using data-adaptive (machine learning) estimation. We will outline a few such possible approaches (including one based on targeted minimum loss-based estimation), compare their properties in a simulation study, and illustrate their use on a motivating application using data from the UCLEB consortium investigating the metabolomic pathways through which six common cardiovascular SNPs act.

EO734 Room MAL B18 GRAPHICAL MARKOV MODELS IV Chair: Helene Massam

E0848: **Causal inference in partially linear structural equation models with Gaussian noise**
Presenter: **Dominik Rothenhaeusler**, ETH Zurich, Switzerland
Co-authors: Peter Buehlmann, Jan Ernest
The identifiability and estimation of partially linear additive structural equation models with Gaussian noise (PLSEMs) is considered. Existing identifiability results in the framework of additive SEMs with Gaussian noise are limited to linear and nonlinear SEMs, which can be considered as special cases of PLSEMs with vanishing non-parametric or parametric part, respectively. We close the wide gap between these two special cases by providing a comprehensive theory of the identifiability of PLSEMs by means of (A) a graphical, (B) a transformational, (C) a functional and (D) a causal ordering characterization of PLSEMs that generate a given distribution P. In particular, the characterizations (C) and (D) answer the fundamental question to which extent nonlinear functions in additive SEMs with Gaussian noise restrict the set of potential causal models and hence influence the identifiability.

E0793: **Bayesian Dirichlet Bayesian network scores and the maximum entropy principle**
Presenter: **Marco Scutari**, University of Oxford, United Kingdom
A classic approach for learning Bayesian networks from data is to select the maximum a posteriori (MAP) network. In the case of discrete Bayesian networks, the MAP network is selected by maximising one of several possible Bayesian Dirichlet (BD) scores, the most famous of which is the Bayesian Dirichlet equivalent uniform (BDeu) score. Since the number of possible networks to choose from grows more than exponentially in the number of variables, the uniform prior associated with BDeu makes structure learning computationally feasible, and does not require the elicitation of prior knowledge from experts. We will discuss the impact of this uniform prior on structure learning from an information theoretic perspective, showing how BDeu may violate the maximum entropy principle when applied to sparse data. On the other hand, a previous BDs score arises from a piece-wise prior and it does not appear to violate the maximum entropy principle, even though it is asymptotically equivalent to BDeu.

E1276: **On the attractive properties for estimating and generating distributions of special palindromic Ising models**
Presenter: **Nanny Wermuth**, Chalmers University of Technology, Sweden
Palindromic Ising models are binary quadratic exponential distributions for marginally symmetric binary variables. Their joint probabilities have an additional, special symmetry since these probabilities listed in any lexicographic order remain unchanged after fully reversing its listed elements. The term palindromic stems from linguistics, where it was coined for sequences of characters which read forward and backward in the same way, like in the sentence 'step on no pets'. Such binary distributions have been used first in statistical physics for special types of ferromagnetism, called there zero-mean Ising models. We concentrate on attractive properties for estimating and generating their distributions when all nodes in their undirected, simple and finite graphs can be ordered and interpreted to have at most two parent nodes.

E0705: **Identification with graphical models for time-dependent data**
Presenter: **Vanessa Didelez**, Leibniz Institute for Prevention Research and Epidemiology - BIPS, University of Bremen, Germany
Time-dependent data, e.g. longitudinal data or event histories, form the basis of many investigations. They are typically concerned with the effects of early exposures or sequential treatments on later / repeated outcomes. Some of the issues encountered in the analyses of time-dependent data include time-varying confounding, irregular observation times, drop-out and censoring. These are problems as they may render the target

parameters of interest unidentifiable, e.g. due to non-ignorable drop-out. Different types of graphical models for time-dependent data will be reviewed. We then show how these can be used to characterise situations where target parameters are identified from the available data. A notion central to this characterisation is that of stability. It essentially demands that certain aspects of the underlying joint distribution be equal across regimes of interest, e.g. in the observational regime with irregular observation times and a (possibly hypothetical) experimental regime where a fixed schedule is enforced. We will illustrate how identifiability can be based on this notion of stability in very different contexts, e.g. the identification of causal effects in survival data or the ignorability of the timing of observations in longitudinal studies.

EO508 Room MAL B20 MODELING AND DIMENSION REDUCTION OF EXTREME EVENTS IN HIGH DIMENSIONS Chair: Raphael Huser

E0257: **Principal component decomposition and completely positive decomposition of dependence for multivariate extremes**
Presenter: **Emeric Thibaud**, EPFL, Switzerland
Co-authors: Daniel Cooley
Multivariate regular variation is a framework which is useful for describing tail dependence and estimating probabilities of multivariate extreme events. Dependence for regularly-varying random vectors is described by the angular measure. In large dimensions, this measure is difficult to estimate. Inspired by principal component analysis (PCA) in the non-extreme setting, we propose two decompositions of a matrix which summarizes pairwise tail dependence in a regularly-varying random vector. The first decomposition is useful to understand the largest modes of dependence as is done with traditional PCA. The second decomposition is useful for calculating probabilities of extreme regions and for simulation. We illustrate methods with an application to daily precipitation measurements at 44 stations in Switzerland.

E0604: **Similarity-based clustering of extreme losses from the London stock exchange**
Presenter: **Miguel de Carvalho**, The University of Edinburgh / CEAUL, University of Lisbon, United Kingdom
Co-authors: Raphael Huser, Rodrigo Rubio
Rigorous analysis of the magnitude and the dynamics of extreme losses in a stock market is essential for institutional investors, professional money managers, and traders. We develop statistical methods of similarity-based clustering for heteroscedastic extremes, which allow us to group stocks according to their extreme-value index and scedasis function, i.e., the magnitude and dynamics of their extreme losses, respectively. Clustering is performed in a product-space and a tuning parameter is used to control whether more emphasis should be put on the latter or the former. This provides a practical tool to gain more insight into stocks synchronization during periods of stress, and can thus be practically useful for risk management. The analysis reveals an interesting mismatch between the magnitude and dynamics of extreme losses on the London Stock Exchange and the corresponding economic sectors of the affected stocks.

E0765: **Peaks-over-threshold inference for spatio-temporal processes, with an application to European windstorms**
Presenter: **Raphael de Fondeville**, EPFL, Switzerland
Co-authors: Anthony Davison
Classical spatio-temporal models for extremes rely on block maxima, but this approach is limited by computational considerations to a few dozen variables. In order to get a better understanding of extremal dependence and reduce model uncertainties, exploitation of gridded datasets, for example from climate models, is necessary. $\mathcal{R}$-Pareto processes based on a peaks-over-threshold approach, use single extreme events, generalize the notion of exceedance, and have relatively simple mathematical expressions. For spatio-temporal modelling, we focus on the Brown–Resnick model, which relies on classical Gaussian models widely used in applications. An efficient algorithm for censored likelihood allows us to perform inference with hundreds of locations. For higher dimensions and generalized risk functionals, we develop an estimator based on the gradient score, whose numerical complexity is similar to likelihood-based inference methods for Gaussian fields. We develop a spatio-temporal model for extreme winter storms over Europe and apply our method on 'three-second wind gusts' from the reanalysis dataset ERA-Interim, which covers the period 1979–2016. We can then use the model to generate new extreme storms with previously unobserved intensities and spatial pattern.

E1088: **Bayesian model averaging over tree-based dependence structures for multivariate extremes**
Presenter: **Sabrina Vettori**, KAUST, Saudi Arabia
Co-authors: Raphael Huser, Johan Segers, Marc Genton
Describing the complex dependence structure of multivariate extremes is particularly challenging. To tackle this issue we develop a novel statistical algorithm that describes extremal dependence taking advantage of the inherent hierarchical dependence structure of the max-stable nested logistic distribution and that identifies possible clusters of extreme variables using reversible jump Markov chain Monte Carlo techniques. Parsimonious representations are achieved when clusters of extreme variables are found to be completely independent. Moreover, we significantly decrease the computational complexity of full likelihood inference by deriving a recursive formula for the nested logistic model likelihood. The algorithm performance is verified through extensive simulation experiments which also compare different likelihood procedures. The new methodology is used to investigate the dependence relationships between extreme concentration of multiple pollutants in California and how these pollutants are related to extreme weather conditions. Overall, we show that our approach allows for the representation of complex extremal dependence structures and has valid applications in multivariate data analysis, such as air pollution monitoring, where it can guide policy making.

EO154 Room MAL B30 SURVIVAL ANALYSIS WITH DEPENDENT ENDPOINTS Chair: Takeshi Emura

E0364: **The analysis of diabetic retinopathy data: A conditional copula approach**
Presenter: **Candida Geerdens**, Hasselt University, Belgium
Co-authors: Elif Acar, Paul Janssen
In many studies the response of interest is time to a predefined event (e.g. time to blindness). Such event times are often subject to right-censoring. Further, event times can be grouped into clusters and therefore be correlated (e.g. time to blindness in both eyes). Copulas provide a popular tool to account for the within-cluster association in time-to-event data. However, a further complexity arises when the study contains not only the observed event times, but also a covariate. The question is then whether or not the covariate has an effect on (1) the time to event and/or (2) the within-cluster association. We propose, based on a conditional copula model, an estimation and a testing strategy to infer the impact of a continuous cluster-level covariate on the association in clustered right-censored event time data. A local likelihood approach is used to estimate the functional form of the copula parameter and a generalized likelihood ratio test is described to assess its constancy. Data on diabetic retinopathy (blindness) are used as motivation and illustration. We investigate the association between the time to blindness in both eyes correcting for a possible effect of age at onset of diabetes.

E0374: **Measure, model and estimation on the dependence structure of bivariate recurrent event processes**
Presenter: **Jing Ning**, The University of Texas MD Anderson Cancer Center, United States
Bivariate or multivariate recurrent event processes are often encountered in longitudinal studies in which more than one type of event is of interest. There has been much research on regression analysis for such data, but little has been done to measure and model the dependence between recurrent event processes. We propose a time-dependent measure, termed the rate ratio, to assess the local dependence between two types of recurrent event processes, and then extend it to allow for covariate adjustments. A two-level semiparametric regression model is proposed for jointly modeling the

frequency and dependence of bivariate recurrent events: the first level is a proportional rates model for the marginal rates, and the second level is a proportional rate ratio model for the dependence structure. The proposed models and methods are illustrated by a soft tissue sarcoma study to examine the effects of the initial treatments on the marginal frequencies of local/distant sarcoma recurrence and dependence structure between two types of recurrence.

E0983: **Investigation and interpretation of the correlation structure in udder infection times by nested Archimedean copulas**
Presenter: **Roel Braekers**, Hasselt University, Belgium
Co-authors: Leen Prenen, Luc Duchateau
The correlation structure which is often imposed on clustered multivariate time to event data, is in most cases of a simple nature. For example, in the shared frailty model, all pairwise correlations between event times in a cluster are taken the same. In modelling the infection times of the four udder quarters clustered within a cow, more complex correlation structures are possibly required. And if, such more complex correlation structures give more insight into the infection process and its spread over the different udder quarters of the cow. We choose a copula approach to study more complex correlation structures in clustered infection times. Hereby, we are able to model the marginal distributions separately from the association parameters, leaving them unaffected by the imposed association structure between the clustered event times. We use both Archimedean and nested Archimedean copula functions to model the associations. After introducing the different copula models, we compare them using likelihood ratio tests and explore the association structures by conditional probabilities. Afterwards we use simulations to validate the size and power of the different likelihood ratio tests used to discriminate between the copula models. Furthermore we simulate from different copula families to look at the robustness of the association estimates when the association structure is misspecified.

E0573: **Dynamic prediction for time-to-death under the joint frailty-copula model**
Presenter: **Takeshi Emura**, National Central University, Taiwan
A copula-based approach is considered to construct dynamic prediction formulas of predicting time-to-death for a patient. The suggested prediction formulas incorporate the genetic and clinical covariates collected on the patient entry as well as time-to-tumour progression (TTP) available after the entry. We first review the framework of dynamic prediction by introducing prediction formulas, such as the conditional failure function and conditional hazard function. We then demonstrate how the parameters in the prediction formulas are estimated by fitting semi-competing risks data to a model. The joint frailty-copula model is employed for the marginal Cox regression models on TTP and time-to-death as well as for the copula model between the two endpoints. For illustration, we derive dynamic prediction formulas to predict time-to-death for ovarian cancer patients.

EO603 Room MAL B33 FUNCTIONAL DATA Chair: Philip Reiss

E0905: **Penalized covariance smoothing and its impact on functional principal component analysis**
Presenter: **Philip Reiss**, University of Haifa, Israel
Functional principal component analysis often proceeds by estimating the eigenfunctions of the covariance operator after smoothing an initial covariance function estimate. Penalized spline approaches to the covariance smoothing step have recently become popular. But the chosen penalization strategy can influence the final estimates to a degree that has generally been overlooked. In some cases, injudicious penalization can yield the false conclusion that a single eigendirection accounts for virtually all the variance. When the data represent developmental processes, such a conclusion may understate the diversity of developmental trajectories found in the population. This problem, and proposed solutions, are illustrated with data from a magnetic resonance imaging study of human cerebral cortex development.

E1129: **Recent advances in elastic functional data analysis**
Presenter: **Anuj Srivastava**, Florida State University, United States
One of the most important challenges in functional data analysis (FDA) is the the phase variability, the lack of registration in the amplitudes of given functions. Elastic functional data analysis (EFDA) is a branch of FDA that deals with recognizes and deals with this problem directly. While earlier works in EFDA focused on separation of phase and amplitude components, i.e. registration of functions using time warping, and their individual modeling using functional PCA (FPCA), the more recent techniques focus on combining phase-amplitude separation with other statistical techniques. We will summarize progress in EFDA in the following topics: (1) Elastic FPCA: one performs PCA while seeking function registration at the same time. (2) Elastic Functional Regression Models: This problem involves using functional predictors where response variable is scalar, and functional linear regression is a commonly used model. In the elastic approach, one either removes phase components from the predictors during estimation of model parameters, or separates them to form additional predictors. We will describe a single index model that uses elastic functional predictors and a polynomial index term to result in a powerful regression model. (3) If time permits, we will also list some recent advances in shape analysis of complex objects using elastic shape metrics.

E0709: **Simultaneous confidence bands and testing for functional data using the GKF with Applications to DTI fibers**
Presenter: **Fabian Telschow**, University of California San Diego, United States
Co-authors: Armin Schwartzman
The use of the Gaussian Kinematic formula (GKF) is explored for constructing simultaneous confidence bands (SCBs) of the population mean curve in densely observed functional data. The GKF is a non-asymptotic formula – hence valid also for small sample sizes –, which is exploited to estimate thresholds of the maximum of a pointwise t-statistic yielding SCBs. Although the GKF relies on smooth Gaussian processes, we show that – having a CLT for the estimator of the mean curve – our approach produces SCBs with asymptotically precise covering rates even under non-Gaussian errors and observation noise. Moreover, in order to avoid bandwidth choices in practice, we borrow ideas from scale spaces, and generalize the above methodology to obtain SCBs of the smoothed population means valid simultaneously across many bandwidths. The performance of this method is compared to state of the art procedures with a focus on small sample sizes ($N \approx 10-30$), where under the Gaussian paradigm our method gives precise covering rates and outperforms its competitors including bootstrap methods. Additionally, we provide an application to a DTI fiber data set of 15 healthy controls and 15 patients. The method is applied to find and localize significant differences between the mean curves of the two groups.

E0761: **Registration for exponential family functional data**
Presenter: **Julia Wrobel**, Columbia University, United States
Co-authors: Jeff Goldsmith
The problem of aligning curves from exponential family distributions is considered. The approach is based on the combination of alignment and functional principal components analysis, and is facilitated by recent extensions of FPCA to non-Gaussian settings. The motivation comes from the physical activity using accelerometers, wearable devices that provide around-the-clock monitoring of activity and produce non-Gaussian measurements. We apply the proposed methods to activity counts using a binary active vs inactive indicator using a binomial distribution. After alignment, the trajectories show clear peaks of activity in the morning and afternoon with a dip in the middle of the day.

EO059 Room MAL B34 FUNCTIONAL DATA ANALYSIS Chair: Jane-Ling Wang

E0225: **A simple method to construct confidence bands in functional linear regression**
Presenter: **Kengo Kato**, University of Tokyo, Japan
Co-authors: Masaaki Imaizumi
The aim is to develop a simple method to construct confidence bands, centered at a principal component analysis (PCA) based estimator, for the slope function in a functional linear regression model with a scalar response variable and a functional predictor variable. The PCA-based estimator is a series estimator with estimated basis functions, and so construction of valid confidence bands for it is a non-trivial challenge. We propose a confidence band that aims at covering the slope function at "most" of points with a prespecified probability (level), and prove its asymptotic validity under suitable regularity conditions. Importantly, this is the attempt to derive confidence bands having theoretical justifications for the PCA-based estimator. We also propose a practical method to choose the cut-off level used in PCA-based estimation, and conduct numerical studies to verify the finite sample performance of the proposed confidence band. Finally, we apply our methodology to spectrometric data, and discuss extensions of our methodology to cases where additional vector-valued regressors are present.

E0504: **Dynamical segmentation of a functional data sequence**
Presenter: **Jeng-Min Chiou**, Academia Sinica, Taiwan
Co-authors: Yu-Ting Chen
A two-stage approach is presented to detect multiple changes in the mean functions of a functional data sequence with an application to road traffic segmentation. The optimality of the segmentation is characterized by minimizing the trace of a covariance operator of the random functions restricted to an L_2 subspace. The aim is to estimate the unknown number and positions of the change points that define the segments. The method first searches for change points recursively with a prespecified number using the optimal segmentation criterion, and then ascertain these to be the change points by backward elimination using hypotheses testing on the covariance operators. We present the consistency result of the algorithm in identifying multiple change points. We examine its practical performance through a simulation study under various scenarios of the total number along with the positions of the change points and apply the method to segmentation of freeway traffic conditions.

E0608: **Detection and identification in functional settings**
Presenter: **Antonio Cuevas**, Autonomous University of Madrid, Spain
Co-authors: Ricardo Fraiman
A model is considered in which n random functions X_i are independently observed in such a way that a block of them (whose set of indices C_n is supposed to belong to a given class $\mathcal{C}_n$ of subsets of $\{1,\ldots,n\}$) might have a distribution G different from the (known) distribution F of the remaining observations. In this setting, we consider both, the detection problem of testing $H_0 : F = G$ versus $H_1 : F \neq G$ and the identification problem of estimating (when H_1 is accepted) the set C_n of exceptional indices where the distribution of the X_i is different from that of the "bulk" of the data. These problems of detection and identification of sparse segments arises, for example, in genomic studies. Most literature on the topic is focused on the one-dimensional case, often assuming in addition that the X_i are Gaussian. We tackle the problem for the case of functional data. We show that the Radon-Nikodym (RN) derivatives for the distributions of stochastic processes (especially in the Gaussian case) are a useful tool for these purposes. In short, our proposal is "to take the data" to the real line, using a suitable RN-derivative, and then to use a detection-identification procedure based on the Kolmogorov-Smirnov (or the Cramer-von Mises) statistic.

EO599 Room MAL B35 STATISTICAL INFERENCE FOR FMRI DATA Chair: Armin Schwartzman

E0423: **Ironing out the wrinkles in the cluster failure**
Presenter: **Anders Eklund**, Linkoping university, Sweden
Functional magnetic resonance imaging (fMRI) results in nice brain activity maps, but many of the used statistical methods have not been validated using real data. We recently showed that parametric statistical methods for cluster inference in fMRI can lead to inflated false positive rates, due to the fact that two assumptions for Gaussian random field theory are violated in real data (a Gaussian spatial autocorrelation function, and a constant spatial smoothness in the brain). A permutation test was for two-sample t-tests also shown to result in nominal false positive rates, as the permutation test is not based on these two assumptions. However, for one-sample t-tests, the permutation approach (based on sign flipping) failed for some parameter combinations. We will therefore focus on how to improve the sign flipping test for one-sample t-tests.

E1371: **A Bayesian credible subgroups approach to statistical inference in fMRI**
Presenter: **Mark Fiecas**, University of Minnesota, United States
The multiple testing problem in fMRI will be discussed. Multiple comparisons due to the hundreds of thousands of voxels in the data inflate the family-wise error rate, and the strong correlation between voxels make exacerbate the problem by making naive multiplicity corrections too conservative. We extend the idea of subgroup identification developed for clinical trials to neuroimaging studies, where we translate the problem to finding the voxels whose time course covaries with the stimulus presentation in the experiment. We develop a Bayesian credible subgroups approach using hierarchical linear models that constructs the subgroups of voxels using information about model parameters from the hierarchical model. These subgroups correspond to one of the following possibilities: i) the voxels with strong evidence of association with the stimulus, ii) the voxels with no evidence of association with the stimulus, and iii) the voxels for which there is insufficient evidence of association with the stimulus. These subgroups allow for one to draw conclusions on the effect of the stimulus on each voxel, and when to defer classification to further collection of data. Finally, our approach fully accounts for the multiplicity of the problem.

E1353: **A hypothesis testing view of searchlight pattern analysis (MVPA)**
Presenter: **Jonathan Rosenblatt**, Ben Gurion University of the Negev, Israel
Searchlight Multi-Voxel Pattern Analysis (MVPA) has been tremendously popular in the neuroimaging community since its introduction, about 10 years ago. The idea of fitting a local/scan/searchlight classifier, can also be found in the genetics literature. We will outline a typical MVPA analysis pipeline and cast it as a statistical multivariate hypothesis test so that it may be compared to the mass univariate approach (i.e.- multiple univariate testing). Seen as a multivariate testing problem, we will discuss the implied hypotheses, potential power gains, and computational shortcuts.

E1398: **Error control in fMRI using the (nonstationary) Gaussian kinematic formula**
Presenter: **Armin Schwartzman**, University of California, San Diego, United States
Co-authors: Dan Cheng, Fabian Telschow, Robert Adler
In fMRI, activation maps are typically excursion sets of a regression coefficient map, seen as a smooth Gaussian random field. The Euler Characteristic (EC) heuristic connects the error control to the expected EC, given by the Gaussian kinematic formula (GKF). While the GKF is remarkably simple, equal to a linear function of the Lipschitz-Killing curvatures (LKCs) of the field over the domain, the use of the GKF has been limited because the LKCs are difficult to compute or estimate for nonstationary fields. Given repeated observations of a field, a consistent estimator of the LKCs is proposed as a simple affine function of the observed EC curves. The estimator is easy to implement and shown to be consistent as the number of repeated observations increases, regardless of the underlying covariance structure. The proposed estimator is used to determine

activation in fMRI and establish spatial error control, using central limit theorems to determine the accuracy in the inference.

EO404 Room Bloomsbury STATISTICAL MODELLING IN INSURANCE Chair: Vali Asimit

E1102: **Security posture assessment in conflict areas: The value of spatial-temporal information**
Presenter: **Enrico Biffis**, Imperial College Business School, United Kingdom
Co-authors: Davide Benedetti
The problem of assessing the security posture of organizations operating in conflict areas is considered. We show how spatio-temporal risk models can be used to understand the dynamic and heterogenous nature of attack occurrences and severities, providing insights into the design of the Close Protection security layer. We use a granular dataset on attacks carried out in Iraq during 2007-15 to quantify the economic gains from using granular information on the actual environment organizations operate in. We quantify such gains by comparing the cost of implementing security postures based on conditional vs. unconditional information. Metrics based on expected casualties are also discussed. We provide practical applications of the model by studying in detail four different areas in Iraq presenting different socio-economic characteristics and patterns of attack occurrence. We then look more explicitly at the oil and gas industry and consider a case study based on a medium sized oil field in the Basra region. We find that appropriate use of spatio-temporal information can deliver average security cost savings of around 30% relative to the case of unconditional security postures, and of around 50% relative to the case of security postures driven by overreaction to spikes in conflict activity that have limited bearing for the exposures at stake.

E0630: **Modelling longevity trends and longevity risk**
Presenter: **Valeria D Amato**, University of Salerno, Italy
Co-authors: Steven Haberman
The fact that life is longer in many developed (and developing) countries has a significant financial effect on individuals, governments, social security systems, pension plans and insurance and reinsurance companies. In this context, we will define longevity risk and we will examine the background in terms of historical trends in mortality rates and longevity and consider the financial implications. In order to plan in advance for these changes, we require reliable models that enable the reliable forecasting of future longevity trends and the measurement of uncertainty. We will examine different approaches to the modelling of the trends in the underlying mortality rates as well as the mortality improvement rates. We will present some results from comparative studies of modelling and forecasting. We will finally reflect on the state of the science and look at the new avenues for research that look promising.

E1190: **In-sample forecasting: Continuous chain ladder and extensions**
Presenter: **Munir Hiabu**, Cass Business School, United Kingdom
The standard method to forecast outstanding liabilities in non-life insurance is the Chain Ladder Method. Data is hereby aggregated into a so-called loss triangle. The actuary handling the claims can usually choose from a drop-down menu wether the data should be aggregated in years, quarters, or months. We will argue that a) aggregation should be understood as a smoothing technique balancing bias and variance. b) If individual data is available, forecasts can be improved by replacing aggregation with kernel smoothers. c) In the continuous version with kernel smoothers, calendar time effects and operation time effects can be incorporated - effects classical Chain Ladder cannot handle. Asymptotic results, a data example, and a simulation study will be provided.

E0750: **Sensitivity analysis without repeated model runs**
Presenter: **Andreas Tsanakas**, City, University London, United Kingdom
Co-authors: Pietro Millossovich, Silvana Pesenti
In risk management, an internal model consists of three elements: (i) a random vector of input risk factors, (ii) a real valued aggregation function, and (iii) the output, which is a random variable obtained by applying the aggregation function on the vector of risk factors. Sensitivity analysis often requires evaluation of changes in the distribution of the output, when the distribution of risk factors is varied with reference to a baseline input distribution. Typically, the distribution of the output is determined via simulation methods. When evaluation of the aggregation function is computationally expensive, as is often the case with models used in practice, extensive sensitivity analyses may become impractical. We propose a method for overcoming this difficulty, by approximating changes in the distribution of the output, while working with a single simulated sample from the baseline model. The method requires knowledge or estimation of the aggregation functions gradient. The approximation is exact when risk factors are independent and gives excellent results when risk factors are not independent, as demonstrated by numerical examples.

EO435 Room G3 STOCHASTIC MODELLING OF ENERGY MARKETS Chair: Almut Veraart

E1327: **Pricing wind power futures**
Presenter: **Brenda Lopez Cabrera**, Humboldt Universitat zu Berlin, Germany
Co-authors: Awdesch Melzer, Wolfgang Haerdle
With increasing wind power penetration more and more volatile and weather dependent energy is fed into the German electricity system. Wind power derivatives were introduced to manage the risk of windless days and transfer the risk of unstable revenue streams from wind turbine owners to third parties. These insurance-like contracts allow to hedge the risk of unstable wind power production on exchanges like Nasdaq, European Energy Exchange. The pricing of wind power derivatives has been only theoretically studied either with GARCH(1,1)-CAR(p) or employing wavelets. We present a new method to price weather derivatives with very skewed data incorporating extreme events in modeling seasonal volatility and compare with previos approaches in transformed Gaussian and pure non-Gaussian CARMA(p; q) models. Our results indicate that our transformed Gaussian CARMA(p; q) model is preferred over the non-Gaussian alternative with Levy increments. The calibration of the empirical market price of risk shows typical behavior for futures and forwards in electricity markets.

E0678: **Empirics and analytics for intraday power markets**
Presenter: **Ruediger Kiesel**, University Duisburg-Essen, Germany
An introduction to short-term electricity markets is given. We will start with the relation of day-ahead and intraday prices on the EPEX for deliveries in Germany/Austria. In the sequel we will focus on analyzing the intraday market. We will discuss empirical properties of intraday power markets and point out development in recent years.

E0670: **Risk premia in cash-settled forward contracts**
Presenter: **Nina Lange**, Technical University of Denmark, Denmark
Co-authors: Nikos Nomikos, Jonas Lager
The aim is to investigate the risk premium in cash settled forward contracts on the Baltic Exchange Indices the so-called Forward Freight Agreements in the dry bulk shipping markets. We estimate multiple spot price models using Markov Chain Monte Carlo. Using a structure-preserving measure change, we then calibrate the risk premium of traded FFA contracts. Finally we link the risk premium to explanatory variables like e.g.,

oil prices, demand and supply for shipping and the state of the global economy and compare the risk premium behaviour with other energy risk premia.

E0343: **Mid-term electricity price forecasting using future data**
Presenter: **Florian Ziel**, University of Duisburg-Essen, Germany
Co-authors: Rick Steinert

Due to the liberalization of markets, the change in the energy mix and the surrounding energy laws, electricity research is a dynamically altering field with steadily changing challenges. One challenge is to provide reliable mid-term forecasts despite high variation in the time series of electricity prices. This issue is tackled in a promising and novel approach. By utilizing the high precision of econometric autoregressive models and the expectations of market participants reflected in future prices, we show that the forecasting performance can be vastly increased while maintaining hourly precision. We investigate the day-ahead electricity price of the EPEX Spot for Germany and Austria and setup a model which incorporates the Phelix future of the EEX for Germany and Austria. The model can be considered as an AR24-X model with one distinct model for each hour of the day. We are able to show that future data contains relevant price information for future time periods of the day-ahead electricity price. By implementing a fast and efficient lasso estimation approach we demonstrate that our model can outperform several other models of the literature.

EO240 Room CLO 101 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING III Chair: Michele Guindani

E1373: **Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso**
Presenter: **Sameer Deshpande**, The University of Pennsylvania, United States
Co-authors: Veronika Rockova, Edward George

A Bayesian procedure is proposed for simultaneous variable and covariance selection using continuous spike-and-slab priors in multivariate linear regression models where q possibly correlated responses are regressed onto p predictors. Rather than relying on a stochastic search through the high-dimensional model space, we develop an ECM algorithm similar to the EMVS procedure targeting modal estimates of the matrix of regression coefficients and residual precision matrix. Varying the scale of the continuous spike densities facilitates dynamic posterior exploration and allows us to filter out negligible regression coefficients and partial covariances gradually. Our method is seen to substantially outperform regularization competitors on simulated data. We demonstrate our method with a re-examination of data from a recent observational study of the effect of playing high school football on several later-life cognition, psychological, and socio-economic outcomes.

E1321: **Joint modelling of recurrent events and survival: A Bayesian nonparametric approach**
Presenter: **Maria De Iorio**, UCL, United Kingdom
Co-authors: Paulon Giorgio, Alessandra Guglielmi, Francesca Ieva

Heart failure (HF) is one of the main causes of morbidity, hospitalization and death in the western world and the economic burden associated with HF management is relevant and expected to increase in the future. We consider hospitalization data for heart failure in the most populated Italian Region, Lombardia. Data were extracted from the administrative data warehouse of the regional healthcare system. The main clinical outcome of interest is time to death and research focus is on investigating how recurrent hospitalizations affect the time to event. The main contribution is to develop a joint model for gap times between two consecutive hospitalizations and survival time. The probability models for the gap times and for the survival outcome share a common patient specific frailty term. Using a Bayesian nonparametric prior as the random effects distribution accounts for patient heterogeneity in recurrent event trajectories. Moreover, the joint model allows for dependent censoring of gap times by death or administrative reasons and for the correlations between different gap times for the same individual. It is straightforward to include covariates in the survival and/or recurrence process through the specification of appropriate regression terms. Posterior inference is performed through Markov chain Monte Carlo methods.

E1260: **Multidimensional wavelets with adaptive random partitioning and its application to probabilistic image process**
Presenter: **Meng Li**, Rice University, United States
Co-authors: Li Ma

The aim is to introduce a probabilistic model-based technique called WARP, or wavelets with adaptive random partitioning, with which multidimensional signals can be represented by a mixture of one-dimensional (1D) wavelet decompositions. A probability model, in the form of randomized recursive partitioning, is constructed on the space of wavelet coefficient trees, allowing the decomposition to adapt to the geometric features of the signal. In particular, when combined with the Haar basis, we show that fully probabilistic function estimation can be carried out in closed form using exact recursive belief propagation. We demonstrate that with WARP, even simple 1D Haar wavelets can achieve excellent performance in image denoising via numerical experiments, outperforming state-of-the-art multidimensional wavelet-based methods especially in low signal-to-noise ratio settings.

E0303: **Convex mixture regression for quantitative risk assessment**
Presenter: **Daniele Durante**, Bocconi University, Italy
Co-authors: Antonio Canale, David Dunson

There is a considerable interest in studying how the distribution of an outcome varies with a predictor. We are motivated by environmental applications in which the predictor is the dose of an exposure and the response is a health outcome. A fundamental focus is inference on the dose levels associated with a particular increase in risk relative to a baseline. Current methodologies either dichotomize the continuous response or focus on modeling changes with the dose in the expectation of the outcome. These choices may lead to a loss of information and provide a restrictive characterization of the dose–response relation. We instead propose a class of convex mixture regression models that allow the entire distribution of the outcome to be unknown and changing with the dose. To avoid overfitting, we rely on a flexible characterization of the density at the extreme dose levels, and express the conditional density at each intermediate dose as a convex combination of the extremal densities. This representation generalizes popular dose–response models for binary outcomes and facilitates inference on extra risk functions of interest in quantitative risk assessments with continuous outcomes. We develop simple Markov chain Monte Carlo algorithms for implementation, and propose goodness-of-fit assessments. The benefits of our methods are highlighted in simulations and in a study on the impact of dioxin exposure on preterm birth.

EO358 Room CLO 102 SEMI- AND NONPARAMETRIC FRONTIER ESTIMATION Chair: Paul Wilson

E0422: **Predicting recessions in Italy: A nonparametric discrete choice model for time series**
Presenter: **Camilla Mastromarco**, University of Salento - Lecce, Italy
Co-authors: Leopold Simar, Valentin Zelenyuk
Efficiency frontier is estimated for Italian economy using quarterly data from 1995 to 2016. A flexible nonparametric two-step approach on conditional efficiencies allows us to eliminate the dependence of production inputs/outputs on time. The efficiency measure can be interpreted as output gap and employed as a predictor of economic slowdown. Applying a generalised non-parametric quasi-likelihood method in the context of discrete choice models for time series data, we investigate how the spread variable, constructed as the difference between the 10-year German Treasury bond and 10-year Italian Treasury bond and our estimated efficiency scores predict the recession in Italy. By using a dataset from 1995 - 2014 with quarterly frequency we emphasize the usefulness of this model for the prediction of Italian recessions in case of two explanatory variable (the lagged spread and efficiency scores). Our model involves two continuous predictors, the spread and the efficiency scores and one discrete variable, the lagged dependent variable. We find that this flexible nonparametric approach offers additional insights than the usual linear probit frequently used in the literature in this context.

E1044: **Inference for nonparametric productivity networks: A pseudo-likelihood approach**
Presenter: **Cinzia Daraio**, University of Rome La Sapienza, Italy
Co-authors: Rolf Fare, Shawna Grosskopf, Maria Grazia Izzo, Luca Leuzzi, Giancarlo Ruocco, Moriah Bostian
There is a rich literature on the nonparametric estimation of efficiency (Data Envelopment Analysis) based on networks (Network DEA) which typically analyzes the networks in a descriptive rather than statistical framework. The goal is fill this gap by developing a new more general framework for modeling the production process to include estimation of production functions, information theoretic approaches to econometrics, machine learning and statistical inference from the physics of complex systems. We combine this general model with Georgescu Roegen's model of flows and funds according to a previous production function. The proposed statistical approach is to reconstruct the network's topology for nonparametric frontier models, based on recent pseudo-likelihood techniques, allowing us to infer the network topology in a Bayesian framework. We illustrate this approach with an application to assess the scientific productivity of world countries at a disciplinary level.

E0895: **Statistical inference in nonparametric frontier estimation: Recent developments and dynamic extensions**
Presenter: **Leopold Simar**, Universite Catholique de Louvain, Belgium
Co-authors: Alois Kneip, Paul Wilson
Nonparametric frontier estimation has been popularized by the use of envelopment estimators in the family of FDH/DEA estimators. Once the efficient frontier has been estimated, the efficiency of each firm is evaluated by gauging its distance to the estimated efficient frontier. When general features of the production set is the focus, like mean of groups, shape of the production set (convexity or not), returns to scale assumptions, testing separability, etc. a test statistics has to be provided which is often a function of averages of efficiency scores. Basic results have been obtained to derive central limit theorems for means of efficiency scores, where the inherent bias of the FDH/DEA estimators jeopardize the properties of simple, naive averages. Still it is possible to correct this problem by estimating the bias term. This has been used in various testing situations (equality of means of groups of firms, returns to scale assumptions, convexity, separability with respect to some environmental variables, etc.). New direction try to extend now these ideas to dynamic setups, where the Malmquist index is one of the basic tools used to analyze the evolution over time of the production sets. New theoretical developments allow indeed to extend the previous results to Malmquist indices. New version of central limit theorems are available and some Monte-Carlo experiments analyze the behavior of group means of these Malmquitst indices in finite samples.

E1033: **Nonparametric estimation of Lerner indices for measuring bank competition**
Presenter: **Paul Wilson**, Clemson University, United States
Co-authors: David Wheelock
The Lerner index (L) is a well-established measure of market power at the level of individual firm and is determined by the spread between market-price of a firm's output (P) and its marginal cost (MC) via $L = (P - MC)/P$. In banking studies, firm-specific Lerner indices are typically estimated using the observed ratio of total revenue to total assets to measure output price P, and computing MC from an estimated parametric,translog cost function. Recent studies have called the translog specification for banks' cost function into question. The local likelihood method is used to estimate the cost function for US bank holding companies while allowing for inefficiency. In addition, we add to total assets the credit-equivalent values of off-balance sheet items to obtain a more comprehensive measure of output price. We compare estimates obtained using one or both of these innovations with estimates using traditional approaches to estimate Lerner indices and show that previous studies have mis-stated the level of competition among US banks.

EO055 Room CLO 203 CIRCULAR MODELS AND THEIR STATISTICAL INFERENCE Chair: Toshihiro Abe

E0444: **On estimating finite mixtures of sine-skewed von Mises distributions**
Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan
Co-authors: Toshihiro Abe, Takayuki Shiohama
Mixtures of the von Mises distributions provide the most popular framework for modelling a population with circular and continuous outcomes arising in a variety of subclasses. In contrast, several circular data are not only multimodal, but also locally asymmetric around one or more modes. We consider an algorithm for estimating a mixture of sine-skewed von Mises distributions from likelihood and Bayesian perspectives, respectively. Furthermore, we use a real data set to show that our proposed model has a better fit than the other models. Consistency of the maximum likelihood estimator is also shown under the topology of the quotient space obtained by collapsing the true parameter set into a single point.

E0440: **Bayesian inference for mode preserving distributions on the circle**
Presenter: **Takayuki Shiohama**, Tokyo University of Science, Japan
Co-authors: Toshihiro Abe, Yoichi Miyata
Unimodal skew circular distributions through inverse monotone functions are considered. General properties of the distributions together with skewness measures based on the distribution and density functions are provided. The inverse k-sine-skewed circular distributions are introduced as special cases of this type. More flat-topped and sharply peaked version of the distributions are also given. General results are also provided for maximum likelihood estimation (MLE) of the parameters and Fisher information matrix of the distributions. To calculate Bayes estimates for the model parameters, we introduce the importance sampling estimation algorithms. We also provide approximate Bayes estimates using Lindley's approximation based on both MLE and MAP estimates. Monte Carlo simulations are performed to compare the performance of the Bayes estimates with the classical estimates. Three circular datasets are analyzed for illustrative purposes.

E0506: **Optimal tests for circular reflective symmetry**
Presenter: **Christophe Ley**, Ghent University, Belgium
Co-authors: Jose Ameijeiras-Alonso, Arthur Pewsey, Thomas Verdebout
Parametric and semiparametric tests of circular reflective symmetry about an unknown central direction are developed that are locally and asymptotically optimal in the Le Cam sense against asymmetric k-sine-skewed distributions. Our findings shed new light on and unify existing results on the same topic, providing inter alia the popular Pewsey test of symmetry with so far unknown efficiency properties. The results from Monte Carlo studies comparing the rejection rates of the tests with those of previously proposed tests lead to recommendations regarding the use of the various tests with small- to medium-sized samples. Analyses of data on the directions of cracks in cemented femoral components and the times of gun crimes in Pittsburgh illustrate the proposed methodology.

E1026: **Fitting mixtures of flexible circular distributions with an application to eye tracking data**
Presenter: **Kees Mulder**, Utrecht University, Netherlands
Co-authors: Irene Klugkist, Ingmar Visser, Daan van Renswoude
In eye tracking research, a common goal is to characterize how groups differ in their visual search strategies. One derived measure of interest is the direction in which a saccade (ie. an eye movement) is made. Usually, humans are more likely to make saccades in the left-right directions, as well as up-down directions. Previously, this type of data was modeled with mixtures of von Mises distributions using the 'movMF' package. We improve upon the usual strategy by fitting mixtures of peaked inverse Batschelet distributions. Frequentist inference can proceed through the EM algorithm. Bayesian inference can proceed through MCMC sampling. In this new approach, fewer parameters are needed, while the data is fit more accurately. Interpretation for this new model is also improved. The R package 'flexcircmix' is available to perform the analysis. In an application it is shown that infants have less precision in saccade directions compared to adults.

EO627 Room CLO 204 SMALL AREA ESTIMATION Chair: Domingo Morales

E0934: **Bayesian small area benchmarking**
Presenter: **Serena Arima**, Sapienza University of Rome, Italy
Co-authors: Alessio Farcomeni
Small area model based estimates can differ widely from the direct estimates, especially for areas with very low sample sizes. One potential difficulty with such estimates is that when aggregated, the overall estimate for a larger geographical area may be quite different from the corresponding direct estimate, the latter being usually believed to be quite reliable. The problem can be more severe when the model is misspecified. An overall agreement with the direct estimates at an aggregate level may sometimes be politically necessary to convince the legislators of the utility of small area estimates. One way to avoid this problem is the so-called "benchmarking", that modifies these model-based estimates so that they are equal to the corresponding aggregate estimate. We propose a Bayesian benchmarking approach that embeds the benchmarking constraint in the model. In particular, we use a constrained distribution for the small area means in order to satisfy the benchmarking constraint. In this way, the small area means are not "corrected" a-posteriori, but they are directly sampled from a constrained distribution. Since all the model parameters are jointly estimated as well as the small area means, their estimations take into account the benchmarking constraint. We present our model using simulated data and compare it with other approaches used in literature.

E0605: **Area-level Poisson mixed models for estimating forest fire occurrences**
Presenter: **Maria Jose Lombardia**, Universidade da Coruna, Spain
Co-authors: Miguel Boubeta, Manuel Marey-Perez, Domingo Morales
Area-level Poisson mixed models are good tools for modelling count data at area level. However, its most basic version has several limitations. It does not take into account complex spatial structures or temporal components. A spatiotemporal extension of the basic Poisson mixed model by incorporating a SAR(1) spatial structure and independent time effects is considered. Model fitting based on the method of moments is proposed. The empirical best predictor of the Poisson parameter is obtained and compared against other competitors such as the synthetic or plug-in estimators. These estimators are empirically investigated by several simulation experiments. As accuracy measure of the proposed estimator, the mean squared error is considered and a bootstrap approach is given. Finally, an application to real data of forest fires occurrences in Galicia during 2007-2008 is carried out.

E0955: **An application of unit-level gamma mixed model to SLCS data**
Presenter: **Tomas Hobza**, Czech Technical University in Prague, Czech Republic
Co-authors: Yolanda Marhuenda, Domingo Morales
Empirical best predictors (EBPs) and plug-in estimators of small area indicators under a unit-level gamma mixed model are considered. The maximum likelihood estimates of the model parameters are obtained by maximizing the Laplace approximation to the model log-likelihood. The behavior of the introduced EBPs and plug-in estimators is studied by a Monte-Carlo simulation experiment. A procedure for bootstrap estimation of the mean squared errors is proposed. Finally, an application of the studied EBPs to data from 2013 Spanish living conditions surveys (SLCS) is presented. The target is to compare the presented gamma mixed model for income variable with the nested error regression model applied to log transformation of the income variable.

E0645: **Bootstrap methods for multiple comparison in small areas statistics**
Presenter: **Stefan Sperlich**, Univserity of Geneva, Switzerland
A huge effort has been made in order to estimate the mean squared error for estimators in small areas. An increasingly popular approach for approximating these errors is that of using resampling methods. In the more general literature on mixed effects or multilevel morels there exist already various proposals of different bootstrap procedures. These have also been proposed in order to directly construct prediction intervals for specific estimators in small area estimation. A generally known problem is that these are typically constructed in a way that they exhibit individual (or say average) coverage probability in the sense that for 95 percent PIs we have 5 percent of areas for which the PI does not include the parameter of interest. We do not know anything about joint converage probabilities and can not use them for comparison of two or more areas. In a first step we try to find out which of these bootstrap methods can easily be extended to exhibit a joint or uniform coverage probability.

EO536 Room MAL 402 MODELLING COMPLEX DATA Chair: Garth Tarr

E0592: **Semiparametric observation-driven models for time-series of count**
Presenter: **Thomas Fung**, Macquarie University, Australia
The aim is to show that observation-driven time-series models, such as the popular integer-valued generalized autoregressive conditional heteroskedastic (ingarch) and generalized linear autoregressive and moving average (glarma) frameworks, can be fit without correctly specifying the family of conditional distributions for the responses. Instead, the underlying family of conditional response distributions is treated as an infinite-dimensional parameter to be estimated from the data simultaneously with the usual finite-dimensional parameters. A key feature of this semiparametric approach is that it can automatically handle over and underdispersion, as well as other deviations from parametric models. Numerical studies suggest that both estimation and inferences using the approach can perform as well as correctly-specified parametric models, but can be more robust under model misspecification. Examples are given to illustrate the practical use of the methods.

E0918: **Outlier detection in a complex linear mixed model: An application in plant breeding trial**
Presenter: **Emi Tanaka**, University of Sydney, Australia
Outlier detection is an important preliminary step in the data analysis often conducted through a form of residual analysis. A complex data, such as those that are analysed by linear mixed models, gives rise to distinct levels of residuals and thus offers additional challenges for the development of an outlier detection method. Plant breeding trials are routinely conducted over years and multiple locations with the aim to select the best genotype as parents or commercial release. These so-called multi-environmental trials (MET) is commonly analysed using linear mixed models which may include cubic splines and autoregressive process to account for spatial trends. We present the use of a mean/variance shift outlier model that fits well into the standard framework of linear mixed models, thus can be easily incorporated into a standard linear mixed model software, and is computationally efficient for routine use. We illustrate its application to a simulation based on a set of real wheat yield trials.

E0614: **Semiparametric regression using variational approximations**
Presenter: **Chong You**, University of Nottingham, Ningbo China, China
Co-authors: Francis Hui, Samuel Mueller, Han Lin Shang
Semiparametric regression offers a flexible framework for modeling nonlinear relationships between a response and covariates. A prime example are generalized additive models where for example, spline bases are used with a quadratic penalty to control for overfitting. Estimation and inference is then generally performed either based on penalized likelihood or under a mixed model framework. Penalized likelihood is fast but potentially unstable, and choosing the smoothing parameters needs to be done externally using cross-validation for instance; the mixed model framework tends to be more stable and offers a natural way for choosing the smoothing parameters, but for non-normal responses involves an intractable integral. We present a new framework for semiparametric regression based on variational approximations. The approach possess the stability and natural inference tools of the mixed model framework, while achieving computation times comparable to penalized likelihood. Focusing on GAMs, we derive fully tractable variational likelihoods for some common response types. We present several advantages of the VA framework for inference. We demonstrate consistency of the VA estimates and asymptotic normality for the parametric component.

EO039 Room MAL 414 NEW ADVANCES IN ANALYSIS OF COMPLEX COHORT STUDIES Chair: Hyokyoung Grace Hong

E0433: **Dimension reduction through unsupervised learning on predictors**
Presenter: **Jae Keun Yoo**, Ewha Womans University, Korea, South
A methodology is proposed to estimate the central subspace in sufficient dimension reduction context by utilizing unsupervised learning on a set of predictors. By using the additional information on the predictors, the response is sliced within each class of the predictors. The potential advantage of this double slicing scheme is no requirements of linearity and constant variance conditions in methodological developments, which are normally assumed in most sufficient dimension reduction methodologies. Also, the central subspaces can be exhaustively estimated under mild condition. Numerical studies confirm the theories, and real data analyses are presented.

E1131: **Integrated powered density: Screening ultrahigh-dimensional covariates with survival outcomes**
Presenter: **Hyokyoung Grace Hong**, Michigan State University, United States
Modern biomedical studies have yielded abundant survival data with high-throughput predictors. Variable screening is a crucial first step in analyzing such data, for the purpose of identifying predictive biomarkers, understanding biological mechanisms and making parsimonious predictions. To nonparametrically quantify the relevance of each candidate variable to the survival outcome, we propose integrated powered density (IPOD), which compares the differences in the covariate-stratified distribution functions. This proposed new class of statistics, with a flexible weighting scheme, is general and includes the Kolmogorov statistic as a special case. Moreover, the method does not rely on rigid regression model assumptions and can be easily implemented. We show that our method possesses sure screening properties, and confirm the utility of the proposal with extensive simulation studies. We apply the method to analyze a multiple myeloma study on detecting gene signatures for cancer patients survival.

E1124: **Individualized multi-directional variable selection**
Presenter: **Xiwei Tang**, University of Virginia, United States
Co-authors: Annie Qu
An individualized variable selection approach is proposed to select different relevant variables for different individuals. In contrast to conventional model selection approaches, the key component of the new approach is to construct a separation penalty with multi-directional shrinkages including zero, which facilitates individualized modeling to distinguish strong signals from noisy ones. As a byproduct, the proposed model identifies subgroups among which individuals share similar effects, and thus improves estimation efficiency and personalized prediction accuracy. Another advantage of the proposed model is that it can incorporate within-subject correlation for longitudinal data. We provide a general theoretical foundation under a double-divergence modeling framework where the number of subjects and the number of repeated measurements both go to infinity, and therefore involves high-dimensional individual parameters. In addition, we present the oracle property for the proposed estimator to ensure its optimal large sample property. Simulation studies and an application to HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.

E1105: **Semi-parametric method for non-ignorable missing in longitudinal data using refreshment samples**
Presenter: **Lan Xue**, Oregon State University, United States
Missing data is one of the major methodological problems in longitudinal studies. It not only reduces the sample size, but also can result in biased estimation and inference. It is crucial to correctly understand the missing mechanism and appropriately incorporate it into the estimation and inference procedures. Traditional methods, such as the complete case analysis and imputation methods, are designed to deal with missing data under unverifiable assumptions of MCAR and MAR. The purpose is to identify and estimate missing mechanism parameters under the non-ignorable missing assumption utilizing the refreshment sample. In particular, we propose a semi-parametric method to estimate the missing mechanism parameters by comparing the marginal density estimator using Hirano's two constraints along with additional information from the refreshment sample. Asymptotic properties of semi-parametric estimators are developed. Inference based on bootstrapping is proposed and verified through

simulations.

EO477 Room MAL 416 INFERENCE FOR TIME-VARYING NETWORKS Chair: Tom Snijders

E0893: **Comparing estimation methods for stochastic actor-oriented models**
Presenter: **Viviana Amati**, University of Konstanz, Germany
Co-authors: Felix Schoenenberger, Tom Snijders
Stochastic actor-oriented models are models for dynamic network data, collected by observing a network and an actor-level variable (referred to as 'behaviour') over time in a panel design. The method of moments (MoM) and the maximum likelihood estimation (MLE) have been developed to estimate the parameters of these models. Recently, focusing on dynamics of networks without a behaviour variable, we proposed an estimator based on the generalized method of moments (GMoM) and we described the algorithmic issues that had to be solved to obtain a stable estimation procedure. We apply this estimator to the interdependent dynamics of a network and a behavioural variable, complementing the usual cross-lagged statistics for the MoM by statistics that are contemporaneous combinations of network and behavioral data. We compare the efficiency of the GMoM with respect to the MoM and the MLE. Moreover, we discuss how differences in the estimated coefficients obtained by the three methods may be used to discover model misspecifications.

E0617: **Latent variable modelling of interdependent ego-networks**
Presenter: **Isabella Gollini**, University College Dublin, Ireland
Co-authors: Isabella Gollini, Alberto Caimo, Paolo Campana
Ego-networks consist of a focal node ("ego") and the nodes to whom ego is directly connected to ("alters"). We present a Bayesian latent variable network modelling framework for describing the connectivity structure of interdependent ego-networks (network of ego-networks) by using a latent space approach. This allows us to visualise both the ego-ego and ego-alter relational structures by estimating the positions of all the individuals in a latent "social space". We apply this new methodology using an efficient variational algorithm able to explore the structure of human smuggling network out of Libya (operation Glauco II) consisting of ego-networks based on the wiretaps acquired by the Italian police on 29 suspects during an investigation period lasting from January to October 2014. The statistical challenge with these ego-networks is that the large number of alters (more than 15k people) can potentially be members of several ego-nets.

E0654: **Accounting for uncertainty in stochastic actor-oriented models for dynamic network analysis**
Presenter: **Heather Shappell**, Johns Hopkins University, United States
Co-authors: Eric Kolaczyk
Stochastic Actor-Oriented Models (SAOMs) are designed to capture network dynamics representing a variety of influences on network change in a continuous time Markov chain framework. Developed in the social network setting, these models allow for the testing of hypotheses through the estimation of parameters expressing possible influences on network change. The current framework assumes the observed network edges are free of type I and type II error. However, this is often an unrealistic assumption. We propose a Hidden Markov Model based approach to estimate the error rates and the parameters in the SAOM model. The modeling approach consists of two components: 1) the latent model, which assumes that the unobserved, true networks evolve according to a Markov process as they did in the original SAOM framework; and 2) the measurement model, which describes the conditional distribution of the observed networks given the true networks. An expectation-maximization (EM) algorithm has been developed for estimation, with the incorporation of a particle filtering based sampling scheme due to the enormity of the state space. We present results from a simulation study that demonstrates our method offers great improvement in the accuracy of parameter estimates when compared to the nave approach of just fitting a SAOM. We also apply our method on functional brain networks inferred from electroencephalogram (EEG) data.

E0985: **Hierarchical multilevel analysis of network dynamics: Bayesian estimation and prior sensitivity**
Presenter: **Tom Snijders**, University of Groningen, University of Oxford, Netherlands
Co-authors: Johan Koskinen
Multilevel longitudinal network data sets, by which are meant longitudinal network data sets that were collected according to the same design in multiple, disconnected groups, are starting to be available more and more. This offers new possibilities for generalization and requires new methods of analysis. A multilevel version of the stochastic actor-oriented model (SAOM) is presented. In a hierarchical model for such data structures, there are two sets of parameters: parameters at the highest level, the population of groups, which may be called the population parameters; and parameters at the group (or network) level. For the joint analysis of group- and population-level, a fully Bayesian approach is followed in which the network in each group evolves according to a SAOM; the groupwise parameters are drawn from a multivariate normal distribution; and the parameters of this multivariate normal have a conjugate prior distribution. A special case of the prior is that some of the variances of the multivariate normal are 0. In many applications the number of groups is rather small, so that sensitivity for the prior is an issue. Prior sensitivity is compared for various weakly informative and non-informative priors.

EO577 Room MAL 421 ROBUST RANKING, VOTING, AND COMPARISONS Chair: Jochen Einbeck

E1328: **Model-selection for ranking data models**
Presenter: **Thomas Brendan Murphy**, University College Dublin, Ireland
Co-authors: Lucy Small
Mallows model is commonly used to model rank data, most often using a form based on Kendall's distance. Generalised Mallows Models (GMM) were developed more recently which expand on the original model. We fit mixtures of these GMMs to rank data, using a Metropolis within Gibbs algorithm. We consider the issue of selecting the most appropriate GMM in practice using a number of different approaches. Applications to simulated data, voting data and wine tasting data are considered.

E0751: **Identifying influential observations in complex Bayesian mediation models with ordinal outcome**
Presenter: **Sarka Rusa**, Charles University, Czech Republic
Co-authors: Arnost Komarek, Emmanuel Lesaffre, Luk Bruyneel
Although increasingly complicated (moderated) mediation models are being employed in practice, most of existing mediation literature has not dealt with model diagnostics. We propose a Bayesian approach to the detection of influential observations (or sets of observations). Importance sampling with weights which take advantage of the dependence structure in mediation models is utilized in order to estimate the case-deleted posterior means of the parameters. The method is applied to the ordinal measurements of patients' willingness to recommend hospitals collected on patients in a large European study to answer the research question whether the outcome depends on recorded system-level features in the organization of nursing care, and whether the related effect is mediated by two measurements of nursing care left undone and possibly moderated by nurse education.

E1086: **Bayesian mixture of extended Plackett-Luce models for the analysis of preference rankings**
Presenter: **Cristina Mollica**, Sapienza Universita di Roma, Italy
Co-authors: Luca Tardella

Choice behavior and preferences typically involve numerous and subjective aspects that are difficult to be identified and quantified. For this reason their exploration is frequently conducted through the collection of ordinal evidence in the form of ranking data. A ranking is an ordered sequence resulting from the comparative evaluation of a given set of items according to a specific criterion. Multistage ranking models, including the popular Plackett-Luce distribution (PL), rely on the assumption that the ranking process is performed sequentially, by assigning the positions from the top to the bottom one (forward order). A recent contribution to the ranking literature relaxed this assumption with the addition of the reference order parameter, yielding the novel Extended Plackett-Luce model (EPL). Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the frequentist perspective. We propose the Bayesian estimation of the EPL mixture. The Bayesian extension benefits from the data augmentation strategy and the conjugacy of the PL with the Gamma prior distribution, by making use of a Metropolis-Hastings step within the Gibbs Sampling scheme to simulate the discrete reference order parameter. The usefulness of the proposal is illustrated with applications to simulated and real datasets.

E0878: **Assessing uncertainty in posterior intercepts from random effect models**
Presenter: **Jochen Einbeck**, Durham University, United Kingdom
Co-authors: Nick Sofroniou

Posterior intercepts from random effect or variance component models provide an attractive tool for the ranking or 'league tabling' of cluster-level units; at the same time allowing for adjustment to covariates (on the individual or the cluster level) through the inclusion of fixed effects. For instance, one may use such intercepts for the ranking of region–wise mortality rates (where the crude, regional rates are often unreliable due to small observed counts) or for the construction of educational league tables from complex sample surveys. However, it is essential that the variability of these posterior intercepts can be assessed accurately, in order to establish whether two cluster-level units can actually be distinguished in terms of their ranking. This question is examined in the context of the 'nonparametric maximum likelihood' approach to random effect modelling, in comparison to glmer and fixed effect models. The motivating application is the PIAAC literacy survey.

EO148 Room MAL 532 RECENT ADVANCES IN QUANTILE REGRESSION Chair: Carlos Lamarche

E0491: **Nonparametric weighted average quantile derivative**
Presenter: **Ying-Ying Lee**, University of California, Irvine, United States

The aim is to estimate the weighted Average Quantile Derivative (AQD), that is, the expected value of the partial derivative of the conditional quantile function (CQF) weighted by a function of the covariates. We consider two weighting functions: (i) a known function chosen by re- searchers; (ii) an estimated density function of the covariates, parallel to a previous density-weighted average mean derivative. The AQD summarizes the average marginal response of the covariates on the CQF and defines a nonparametric quantile regression coefficient. In semi-parametric single-index and partial linear models, the AQD identifies the coefficients up to scale. In nonparametric nonseparable structural models, the AQD conveys an average structural effect. We interpret and estimate the AQD by a weighted average CQF. Including a stochastic trimming function, the proposed two-step estimator is root-n-consistent for the AQD defined on the entire support of the covariates. A key preliminary result is a new Bahadur- type linear representation of the generalized inverse kernel-based CQF estimator uniformly over the covariates in an expanding compact set and over the quantile levels. Our asymptotic normality results can be extended to the Hadamard-differentiable functionals of the conditional or average quantile processes, which can be nonlinear functionals of the distributions.

E1090: **Quantile regression for longitudinal data: A convex clustering approach**
Presenter: **Jiaying Gu**, University of Toronto, Canada
Co-authors: Stanislav Volgushev

A penalization-based approach is proposed to automatic discovery of group structure in quantile regression with panel data. More precisely, we assume that the observed individuals come from a population with an unknown number of types and each type is allowed to have its own coefficients in a linear quantile regression model. In contrast to existing literature, the number of types in the population does not need to be specified in advance. Rather, a merit of the proposed penalization-based approach is that the number of types and group membership is determined in a data-driven fashion. An adaptive method for finding the tuning parameter is also proposed. Consistency of the procedure for estimating group membership as well as asymptotic normality of the resulting quantile regression coefficients within each group are established.

E1179: **Common correlated effects estimation of heterogeneous dynamic panel quantile regression models**
Presenter: **Carlos Lamarche**, University of Kentucky, United States
Co-authors: Matthew Harding, Hashem Pesaran

A quantile regression estimator is proposed for a heterogeneous panel model with lagged dependent variables and interactive effects. The Common Correlated Effects (CCE) approach is adopted. We demonstrate that the extension to the estimation of dynamic quantile regression models is feasible under similar conditions to the ones used in the literature. We establish consistency and derive the asymptotic distribution of the new quantile regression estimator. Monte Carlo studies are carried out to study the small sample behavior of the proposed approach. The evidence shows that the estimator can significantly improve the performance of existing estimators as long as the time series dimension of the panel is large. We present an application to the evaluation of Time-of-Use pricing using a large randomized control trial.

E1247: **Distributed computing for quantile regression: A statistical analysis**
Presenter: **Stanislav Volgushev**, University of Toronto, Canada
Co-authors: Shih-Kang Chao, Guang Cheng

With emergence of new data collection and storage technologies, it has become easy to accumulate extremely large data sets. At the same time, statistical analysis of such data poses serious computational challenges. One common approach to handling the resulting computational burden relies on splitting the complete data set into smaller subsamples and performing computation on each of the subsamples. While such an approach is easy to implement, the theoretical properties of resulting procedures remain largely unclear. We provide a detailed analysis of such a splitting approach to quantile regression and discuss potential applications, including the analysis of panel data models.

EO106 Room MAL 538 CLUSTER ANALYSIS WITH R Chair: Sebastian Fuchs

E0623: **The mclust R package for clustering, classification and density estimation using Gaussian finite mixture models**
Presenter: **Luca Scrucca**, Universita' degli Studi di Perugia, Italy
Co-authors: Michael Fop, Thomas Brendan Murphy, Adrian E Raftery
Finite mixture models are being used increasingly to model a wide variety of random phenomena for clustering, classification and density estimation. mclust is a powerful and popular R package implementing Gaussian finite mixtures with different covariance structures and different numbers of mixture components. An integrated approach is provided, with functions that combine model-based hierarchical clustering, EM for mixture estimation and several tools for model selection. Recent updates have introduced new covariance structures, dimension reduction capabilities for visualisation, model selection criteria, initialisation strategies for the EM algorithm, and bootstrap-based inference, making it a full-featured R package for data analysis via finite mixture modelling.

E0251: **Cross-entropy clustering**
Presenter: **Przemyslaw Spurek**, Jagiellonian University, Poland
Gaussian Mixture Model (GMM) is one of the most popular clustering models implemented in various R packages, such as mclust, Rmixmod, pdfCluster, mixtools, etc. The model focuses on finding the mixture of Gaussians $f = p_1 f_1 + \ldots + p_k f_k$ where $p_1, \ldots, p_k > 0$ and $\sum_i p_i = 1$, which provides an optimal estimation of data set $X \subset \mathbb{R}^N$, measured by the cost function $\mathrm{EM}(f,X) = -\frac{1}{|X|}\sum_{x\in X} \log\left(p_1 f_1(x) + \ldots + p_k f_k(x)\right)$. Its minimization is iteratively performed with use of EM (Expectation Maximization) algorithm. While the expectation step is relatively simple, the maximization step usually needs complicated numerical optimization. We presents R Package CEC, the first open source implementation of a Cross-Entropy Clustering method, which is a fast hybrid between k-means and GMM. Similarly to GMM, CEC searches for Gaussian densities, which minimizes the cost function $\mathrm{CEC}(f,X) = -\frac{1}{|X|}\sum_{x\in X} \log\left(\max(p_1 f_1(x), \ldots, p_k f_k(x))\right)$. Although the difference between the two functions is slight and relies on substituting the sum operation by the maximum, it occurs that the optimization can be realized in a comparable time to k-means. CEC allows to reduce unnecessary clusters on-line.

E1829: **Cross-Clustering: A partial clustering algorithm with automatic estimation of the number of clusters**
Presenter: **Paola Tellaroli**, University of Padova, Italy
Co-authors: Michele Donato, Marco Bazzi, Alessandra Brazzale, Sorin Draghici
Four of the most common limitations of the many available clustering methods are: i) the lack of a proper strategy to deal with outliers; ii) the need for a good a priori estimate of the number of clusters; iii) the lack of a method able to detect when partitioning of a specific data set is not appropriate; and iv) the dependence of the result on the initialization. We propose Cross-clustering (CC), a partial clustering algorithm that overcomes these four limitations by combining the principles of two well established hierarchical clustering algorithms: Wards minimum variance and Complete-linkage. We validated CC by comparing it with a number of existing clustering methods, including Wards and Complete-linkage. We show on both simulated and real datasets, that CC performs better than the other methods in terms of: identification of the correct number of clusters, identification of outliers, and determination of real cluster memberships. We used CC to cluster samples in order to identify disease subtypes and on gene profiles, in order to determine groups of genes with the same behavior. Results obtained on a non-biological dataset show that the method is general enough to be successfully used in such diverse applications. New results on a modified version of the CC algorithm show that it could be suitable also for identify elongated clusters. The algorithm has been implemented in the statistical language R and is freely available from the CRAN repository.

E0773: **Clustering of variables around latent variables, with the R package ClustVarLV**
Presenter: **Evelyne Vigneau**, National College of veterinary medicine, food science and engineering, France
Co-authors: Veronique Cariou, El Mostafa Qannari
The aim of clustering variables is to group a set of variables into homogeneous and distinct clusters and to identify the underlying structure of the data. In exploratory data analysis, such an approach may be very useful for interpreting complex problems, as it provides a dimension reduction. The clustering of variables around latent variables (CLV) approach, implemented in the ClustVarLV R package, aims at clustering numerical variables along with summarizing each group of variables by a latent component. Directional or local clusters of variables may be defined according to the type of the linear link to be investigated. Moreover, the latent variable associated with the clusters may be constrained to be linear combinations of external information. A new extension has also been implemented within ClustVarLV for the clustering of variables while setting aside atypical ones. In this scope, two strategies have been proposed: the first one consists in introducing an additional group of variables whereas, the second one consists in determining sparse components. This latter aspect will specifically be illustrated on the basis of real case studies.

EO037 Room MAL 539 DOUBLY STOCHASTIC COUNTING PROCESSES Chair: Paula Bouzas

E0587: **Different approaches for testing the independence between point processes**
Presenter: **Ana C Cebrian**, University of Zaragoza, Spain
Many real problems involve two or more point processes and require a description of their dependence structure, for example the timing of the trades and mid-quote changes in Stock Exchange or the occurrence of climate extremes at different spatial locations. These processes have to be studied in a multivariate point process framework and an important issue to consider is the independence between the marginal processes. Three families of tests to check the independence between point processes, conditionally on their marginal structure, are presented: conditional tests for Poisson processes, tests based on the close point sets and tests based on the J and K functions. Each test is based on different assumptions, and is adequate for different type pf processes: Poisson processes, homogeneous point processes, nonhomogeneous processes with a parametric model, etc. All together cover a wide range of situations and dependence structures. A comparative study of the size and power of the tests is carried out. Future work includes the extension of some of the suggested tests to the case of doubly stochastic point processes.

E0651: **Spatial cluster point processes with marks depending on clusters**
Presenter: **Ute Hahn**, University of Aarhus, Denmark
Cluster processes are obtained as the superposition of finite point processes, the daughters, centred in the points of a parent process. We will introduce and discuss a model where the daughter processes are marked point processes, with i.i.d. ground process, but with mark distribution that is different for different clusters and has relatively narrow bounded support. The overall mark distribution is then a mixture of the cluster mark distributions; therefore the marked cluster process is also clustered in mark space. For this model, it can be shown that the pair correlation function of the parent process can be retrieved, up to a convolution, from the pair correlation function and the mark correlation function of the cluster process. The motivation comes from a data set regarding ultra microscopy, where points represent localizations of single molecules. The points are obtained from fluorescence photons that are registered in a video, and thus they are equipped with time marks. Usually, these data sets are only analysed as point processes in space, but the extra information given by the time marks provides interesting insight into the generation of the data.

E0842: **Compound Cox processes applied to extreme meteorological events**
Presenter: **Nuria Ruiz-Fuentes**, University of Jaen, Spain
Co-authors: Paula Bouzas, Carmen Montes-Gijon

Given the enormous socioeconomic and environmental impact of extreme meteorological events, studies that provide predictive data in this area have become increasingly prevalent in recent decades. A series of this type of events, such as extreme values of temperatures (maximum and/or minimum), high levels of ultraviolet radiation or precipitation in a climate zone, may be modeled as a point process or as the resulting counting process; the compound Cox process is proposed for this purpose. Examining available data from different climate zones from 1991 to 2011, their intensity processes were estimated by means of an ad hoc functional principal components model. Having obtained the estimated models, the counting processes can be forecast (i.e. their mean, number of points and the probability of having a new event in a chosen future interval of time within a year). In addition, using a goodness-of-fit test, a new sample path collected in any climate zone can be tested to ascertain if it follows the same stochastic model; thus, the hypothesis test helps to discriminate between climate zones.

E1010: **Phase type process**
Presenter: **Paula Bouzas**, University of Granada, Spain
Co-authors: Nuria Ruiz-Fuentes, Carmen Montes-Gijon

Non-negative valued stochastic processes are often observed when studying real phenomena; consequently, modeling these processes is an interesting issue. Linear combinations of exponentials, gamma or Weibull distributions are commonly encountered in the literature in order to model them. Based on the fact that the most general model for a non-negative valued random variable is the phase type distribution, a new non-negative stochastic model is proposed. The phase type process is defined as a doubly stochastic process whose marginals are phase type distributions with random parameters. Survival, reliability and queuing theory or counting processes are fields in which non-negative stochastic processes play an important role. As an example, a compound Cox process whose intensity is a phase type process is considered. This counting process emerges as especially flexible and general. After this description, a simulation study is presented.

EO579 Room MAL 540 ADVANCES IN JOINT MODELLING Chair: Elisabeth Waldmann

E0821: **Joint modelling of longitudinal and discrete time-to-event data**
Presenter: **Jessica Barrett**, MRC Biostatistics Unit, United Kingdom
Co-authors: Peter Diggle, Robin Henderson, David Taylor-Robinson

Maximum likelihood estimation of joint models with shared latent random effects typically involves numerical integration of the likelihood over the random effects distribution or use of an expectation-maximisation algorithm. We present an estimation method for joint modelling of a longitudinal outcome and a discrete time-to-event outcome. By writing the joint likelihood in the form of a skew-normal distribution, we are able to analytically integrate over the random effects distribution and express it as a multivariate normal cumulative distribution function. This method allows more efficient estimation of joint models with more complex random effects structures, including higher-dimensional random effects. We illustrate our methods with an application to data from the UK Cystic Fibrosis Registry.

E0868: **Extending joint models in terms of boosting algorithms**
Presenter: **Colin Griesbach**, FAU Erlangen-Nuernberg, Germany
Co-authors: Elisabeth Waldmann, Andreas Mayr

Joint models turned out to be a powerful approach to analysing data where event times are measured alongside a longitudinal outcome. The idea is to combine a longitudinal and a survival model via a shared predictor used in both original models while a parameter quantifies their relation. To fit a basic joint model efficiently, a gradient boosting algorithm has been presented based on statistical boosting methods for longitudinal data in multiple dimensions. The aim is to extend that algorithm by incorporating a predictor solely for survival data, hence a set of covariates, which are independent of the longitudinal structure, is added.

E0987: **Flexible two-stage model proposal for multivariate longitudinal and survival data using spline smoothing**
Presenter: **Ipek Guler**, CiMUS, University of Santiago de Compostela, Spain
Co-authors: Christel Faes, Carmen Cadarso Suarez, Francisco Gude

In many biomedical studies, it is common to follow up subjects repeatedly. These follow-up studies typically produce different types of outcomes including both longitudinal biomarkers and time-to-event outcomes. Often, the interest is on assessing the relationship between the longitudinal and the time-to-event processes. Joint modelling approaches of longitudinal and survival data is an appropriate way to study such relationship. Existent joint models are mostly concentrated on a single longitudinal and survival process. However, many studies collect several longitudinal biomarkers and instead of selecting a single longitudinal biomarker we want to study the relationship between all these biomarkers and the survival outcome. Additionally, flexible regression techniques may be necessary for the non-linear longitudinal and survival trends. For instance, Orthotopic Liver Trasplantation data includes glucose and daily insulin therapy which could predict the risk for death of the patients who underwent transplantation. Both longitudinal trends and the risk for death show non-linear profiles. The joint modelling approaches in a frequentist framework are difficult to implement when the number of longitudinal biomarkers is large or when flexible techniques are needed. For this aim we propose a two-stage based modelling approach for modelling of non-linear multivariate longitudinal and non-linear survival data using spline smoothing.

E1493: **A comparison of different R routines for joint modelling**
Presenter: **Anja Rappl**, Friedrich-Alexander Universitaet Erlangen-Nuernberg, Germany
Co-authors: Elisabeth Waldmann

When confronted with data that simultaneously captures outcomes of longitudinal and time-to-event processes it is advisable to model this data jointly. With the rising popularity of these joint models in recent years various `R` packages were developed incorporating the current state of research to facilitate analyses. However, the packages differ in model formulation, estimation techniques and the extent of estimated parameters. For an overview of their capabilities and limitations the packages `JM`, the pioneer amongst packages for joint modelling ,and `JoineRML` allowing for multivariate joint models as well as `JMBoost`, the first package with automated variable selection, will be compared via a simulation study and a real data example. The result may serve as guidelines as to when to use which package as well as further promote the more routine application of joint models.

EO053 Room MAL 541 CHANGE-POINT DETECTION IN TIME SERIES Chair: Herold Dehling

E0753: **Robust change-point estimation in the presence of long-range dependence**
Presenter: **Annika Betken**, Ruhr-Universitat Bochum, Germany

An estimator is considered for the location of a shift in the mean of long-range dependent sequences. The estimation is based on Wilcoxon-type statistics. Consistency and an optimal rate of convergence are established for constant shift heights and and under certain local changes, i.e. for shift heights decreasing to 0 with a certain rate. If suitably standardized, the estimator converges in distribution to a functional of a fractional Brownian motion process. Its asymptotic behavior as well as its finite sample performance are compared to corresponding properties of a CUSUM-type change-point estimator.

E0938: **Limit theorems for multivariate long-range dependent processes**
Presenter: **Marie-Christine Dueker**, Ruhr University Bochum, Germany

Over the last thirty years, long-range dependent stochastic processes become an important instrument for modeling phenomena in econometrics, engineering and hydrology to mention some examples. Moreover, the analysis of high-dimensional time series gets more attention. The study of different types of limit theorems under the assumption of multivariate long-range dependence is considered. This includes the behavior of partial sums of multivariate linear processes, subordinated Gaussian processes and the sample autocovariances. A suitable matrix-valued normalization sequence under the assumption of multivariate LRD, which could be of particular interest for further results in this context, is investigated.

E0866: **Robust change-point estimation**
Presenter: **Carina Gerstenberger**, Ruhr-Universitaet Bochum, Germany

In many applications it cannot be assumed that observed data have a constant mean over time. Therefore, extensive research has been done in testing for change-points and estimation of the change-point location. However, just a few procedures are robust against outliers in the data. We introduce estimators of the location parameter for the change-point in the mean based on U-statistics and establish consistency under short-range dependence. In a simulation study we will see that a suitable choice of the kernel function yields estimators that are robust to outliers and heavy-tailed distributions.

E0815: **Detecting structural breaks via ordinal pattern probabilities: The short- and the long-range dependent framework**
Presenter: **Alexander Schnurr**, University Siegen, Germany
Co-authors: Herold Dehling, Jeannette Woerner, Jannis Buchsteiner

Ordinal patterns describe the order structure of data points over a small time horizon. Using a moving window approach, we reduce the complexity of a time series by analyzing the sequence of ordinal patterns instead of the original data. We present limit theorems for ordinal pattern probabilities and tests for structural breaks in the short-range dependent as well as in the long-range dependent setting. In the long-range dependent case, we investigate the ordinal information of a subordinated Gaussian process with a non-summable autocovariance function. We establish the asymptotic behavior of different estimators for ordinal pattern probabilities by using a multivariate Hermite decomposition. Ordinal pattern dependence is a new way of measuring the degree of dependence between time series. Since it only relies on the ordinal structure of the data, it is robust against monotone transformations and measurement errors. This method has proved to be useful already in the context of hydrological, financial as well as medical data. Using this concept it is possible to analyze whether the dependence structure between two time series changes over time.

EC694 Room MAL 153 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS Chair: Silvia Cagnone

E1481: **A decision-theoretic property of conditional normalized maximum likelihood distribution**
Presenter: **Yoshihiro Hirose**, Hokkaido University, Japan

Distribution prediction is considered, where we observe a data, and estimate the distribution of a future data based on the observation. The estimated distribution is called a prediction distribution. Our target is Conditional Normalized Maximum Likelihood (CNML) distribution. CNML is a generalization of Normalized Maximum Likelihood (NML) distribution. NML is the minimax distribution with respect to the regret. The regret is the difference between a candidate distribution and the best distribution for a virtually-observed value. Similarly, CNML is the minimax distribution with respect to a conditional regret. Three versions of CNMLs corresponding to three types of conditional regrets have been introduced. Based on an observed data, a conditional regret compares a prediction distribution with the best distribution for a virtually-observed value. We are interested in whether CNMLs are admissible or not under some criterion. The admissibility is a basic concern in statistical decision theory. We are also interested in the minimaxity of CNMLs. CNML is originally minimax with respect to the conditional regret. However, it is not clear that it is also minimax under other criterion, e.g., the Kullback-Leibler divergence and conditional regret risks. The result depends on statistical models we assume.

E1544: **Parametric dominance relations for distributions obtained by composition**
Presenter: **Tommaso Lando**, VSB Technical University of Ostrava, Czech Republic
Co-authors: Lucio Bertoli-Barsotti

The so called T-X family provides an interesting new method for generating distributions by composing (possibly parametric) distribution and quantile functions. Such an approach makes it possible to add parameters to an existing model, in order to obtain a new and more flexible one. In this framework, we study the conditions under which distributions obtained through the T-X method can be ranked by a dominance relation, such as the stochastic dominance of first and second degree.

E1623: **Second-order analytic bias reduction for nonlinear panel data models with fixed effects**
Presenter: **Martin Schumann**, TU Dortmund, Germany

One of the most useful features of panel data is that it allows researchers to control for time-invariant individual heterogeneity that is not observed. However, in nonlinear panel data models with fixed effects, the maximum likelihood estimator can be severely biased due to the incidental parameters problem. While in the recent literature methods have been proposed that yield a first-order bias reduction relative to maximum likelihood,simulation results based on short panels suggest the need for higher-order bias reduction in order to further improve the small sample performances of these methods. Explicit expressions for the second-order biases of the profile likelihood and its score are provided. It is further shown that estimation of the first-order bias based on plug-in estimators creates an additional bias that contributes to the second-order bias. Finally, an estimator is constructed that corrects both the first and the second-order bias of the maximum likelihood estimator.

E1489: **Estimation of two-dimensional rate functions using incomplete data**
Presenter: **Micha Mandel**, The Hebrew University of Jerusalem, Israel
Co-authors: Naomi Kaplan-Damary, Yoram Yekutieli, Sarena Wiesner, Yaron Shor

A sample of two-dimensional counting processes having a joint rate function up to a multiplicative constant is considered. The points of each process can be observed only in a process-specific set, leading to truncated data. The aim is to estimate non-parametrically the rate function. We discuss several estimation methods based on mixed logistic and Poisson models and on conditional likelihood. As the approaches are com-

putationally expensive and the number of points in each process is small, case-control sampling is used to reduce the computation effort. The problem is motivated by forensic footwear analysis that is based on randomly acquired characteristics found in the shoe sole. The methods will be demonstrated using a large dataset collected by the Division of Identification and Forensic Science of the Israeli National Police.

EC697 Room MAL 415 CONTRIBUTIONS IN BAYESIAN METHODS Chair: Alexandros Beskos

E1482: **Probabilistic, Bayesian updating of input-output tables: Application to WIOD**
Presenter: **Vladimir Potashnikov**, RANEPA, Russia
Co-authors: Oleg Lugovoy, Andrey Polbin
Developments and applicationa of probabilistic method(s) for updating IO tables are shown. The core of the methodology is a Bayesian framework which combines an information from observed data, additional believes (priors), and related uncertainties into posterior joint distribution of input-output table (IOT) coefficients. The framework can be applied to various IOT problems, including updating, disaggregation, evaluation of uncertainties in the data, and addressing incomplete/missing observations. The flexibility of the methodology is partially based on sampling techniques. We apply modern Monte Carlo Markov Chains (MCMC) methods to explore posterior distribution of IOT coefficients. We also compare results with mainstream methods of updating IOT to investigate its performance. Various indicators of performance and application to various data suggest different results. The overall performance of the method is similar or comparable with mainstream techniques. The main advantage the proposed methodology is an estimation of full profile of joint probability distribution of unknown IOT matrices. The method can be also combined with any other techniques through prior information.

E1616: **Bayesian inference in two-sex branching processes with mutations: ABC approach**
Presenter: **Cristina Gutierrez Perez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Rodrigo Martinez Quintana
A two-sex two-type branching process was recently introduced to analyze the evolution of the numbers of alleles and their mutations of a Y-linked gene. Since genes associated to the Y-chromosome, as well as mutation in there, are of special interest due to their direct relation with male fertility problems or the history of paternal lineages, the interest of how the number of carriers of these genes evolve in a population makes necessary to solve the problem of estimating the main parameters of this model.To this aim, the Approximate Bayesian Computation (ABC) methodology is developed to approximate the posterior distributions of the parameters of such model. For that, it is considered a realistic sampling scheme where the total numbers of females and males (with unrecognized genotypes) are observed in each generation up to determined one. Moreover, from the last generation, the information of the total number of males of different genotypes and those which are mutants among them is added. Finally the accuracy of the procedure using a simulated study is illustrated.

E0336: **Applying Bayesian methods on confidence intervals for quantile estimation**
Presenter: **Margus Pihlak**, Tallinn University of Technology, Estonia
The main idea is the application of the Bayes-bootstrap conversation method on quantiles estimation. This conversation method is based on connecting of Bayesian Inference and parametric bootstrap. Different examples based on raw data will be given.

E1540: **Assessing the stability of response styles by using Bayesian item response modeling**
Presenter: **Kensuke Okada**, Senshu University, Japan
Co-authors: Daiki Hojo, Yusuke Takahashi
A response style is the behavioral tendency of selecting particular categories on the rating scale of questionnaire irrespective of the questionnaire's content. Although several previous studies have investigated the stability of response styles, it is inherently difficult to disentangle the substantive trait of interest from the response styles based on self-rating surveys only. We seek to tackle this problem by using measurements through anchoring vignettes, which consist of brief statements describing the situations of hypothetical individuals. The anchoring vignette data allow us to make use of statistical models to adjust the responses of self-rating assessments. In particular, we propose a new Bayesian model for investigating the longitudinal stability of response styles using anchoring vignettes. This is done by extending the multidimensional item response models to incorporate both the substantive trait and the response style. Markov chain Monte Carlo method is used to carry out Bayesian estimation. Information criterion is used to comparatively evaluate the suitability of the time-stable and time-dependent response style models. The proposed method is illustrated through empirical examples. The results show different tendencies between the substantive traits. We discuss the possible explanations and implications of the findings.

EG016 Room MAL 151 CONTRIBUTIONS IN CONFIDENCE REGIONS Chair: Jelle Goeman

E1636: **On avoiding inconsistencies between confidence intervals and tests for parameters of some discrete distributions**
Presenter: **Jan Klaschka**, Institute of Computer Science of the Czech Academy of Sciences, Czech Republic
Co-authors: Jeno Reiczigel, Mans Thulin
A problem in testing and interval estimation of discrete distribution parameters is addressed: Some confidence interval construction methods, namely those by Sterne (also referred to as the probability based method) and Blaker (also called the combined tails method) suffer from inconsistency between the interval estimates and the corresponding tests. The test rejects, in some cases, hypothesis $\theta = \theta_0$ at significance level α, though parameter value θ_0 lies in the $1-\alpha$ confidence interval. The problem stems from the fact that the set of those θ_0 for which the hypothesis $\theta = \theta_0$ is not rejected, may not be connected, and the gaps have to be filled in order to obtain a confidence interval. A discrepancy then appears when θ_0 lies in a gap. The proposed solution of the problem consists in a modification of the tests (so called *p*-value function unimodalization) that makes the (modified) tests and confidence intervals match perfectly. Programs implementing the test modification for the most frequented settings partly already are, and partly will be soon available on the web.

E1660: **Asymptotic confidence bands in the Spektor-Lord-Willis problem via kernel estimation of intensity derivative**
Presenter: **Zbigniew Szkutnik**, AGH University of Science and Technology, Poland
Co-authors: Bogdan Cmiel, Jakub Wojdyla
The stereological problem of unfolding the distribution of spheres radii from linear sections, known as the Spektor-Lord-Willis problem, is formulated as a Poisson inverse problem and an L2-rate-minimax solution is constructed over some restricted Sobolev classes. The solution is a specialized kernel-type estimator with boundary correction. For the first time for this problem, non-parametric, asymptotic confidence bands for the unfolded function are constructed. Automatic bandwidth selection procedures based on empirical risk minimization are proposed. It is shown that a version of the Goldenshluger-Lepski procedure of bandwidth selection ensures adaptivity of the estimators to the unknown smoothness. The performance of the procedures is demonstrated in a Monte Carlo experiment.

E0738: **Combining confidence intervals: Uncertainty in normed test scores due to test unreliability and sampling variability**
Presenter: **Lieke Voncken**, University of Groningen, Netherlands
Co-authors: Casper Albers, Marieke Timmerman
Test publishers usually provide confidence intervals for normed test scores that reflect the uncertainty due to the unreliability of the tests. The uncertainty due to sampling variability in the norming phase is ignored in practice. To enable a fair positioning of the person under study relative to the norm population, it is important to account for both sources of uncertainty. A flexible method is proposed that combines both types of uncertainty in one confidence interval. This method is applicable in continuous norming and is very flexible in terms of the score distribution, using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework. The performance of the method is assessed in a simulation study. The findings are discussed and the method is illustrated with real norming data.

E1495: **Bootstrap confidence bands for spectral estimation of Levy densities under high-frequency observations**
Presenter: **Daisuke Kurisu**, University of Tokyo, Japan
Co-authors: Kengo Kato
Bootstrap methods are developed to construct uniform confidence bands for nonparametric spectral estimation of Lévy densities under high-frequency observations. We are given n discrete observations at frequency $1/\Delta$, and assume that $\Delta = \Delta_n \to 0$ and $n\Delta \to \infty$ as $n \to \infty$. We employ a spectral estimator of the Lévy density, and develop novel implementations of multiplier and empirical bootstraps to construct confidence bands on a compact set away from the origin. We provide conditions under which the confidence bands are asymptotically valid. We also develop a practical method for bandwidth selection, and conduct simulation studies.

EG012 Room MAL 152 CONTRIBUTIONS IN CONTROL CHARTS Chair: Schalk Human

E1659: **Generally weighted moving average control charts**
Presenter: **Schalk Human**, University of Pretoria, South Africa
Co-authors: Niladri Chakraborty, Balakrishnan Narayanaswamy
Distribution-free control charts gained momentum in recent years as they are more efficient in detecting a shift when there is a lack of information regarding the underlying process distribution. However, a distribution-free control chart for monitoring the process location often requires information on the in-control process median. This is somewhat challenging because, in practice, any information on the location parameter might not be known in advance and estimation of the parameter is therefore required. Parameter estimation from an in-control reference sample typically requires a large number of observations to attain reasonable chart performance when quick detection of a small shift in the location is important. In view of this, a time-weighted control chart, labelled the Generally Weighted Moving Average (GWMA) exceedance (EX) chart (in short GWMA-EX chart), is proposed for detection of a shift in the unknown process location; this chart is based on an exceedance statistic when there is no information available on the process distribution. An extensive performance analysis shows that the proposed GWMA-EX control chart is, in many cases, better than its contenders.

E1357: **Multivariate statistical process control using STATIS method**
Presenter: **Danilo Marcondes Filho**, Universidade Federal do Rio Grande do Sul, Brazil
Co-authors: Luiz Paulo Luna de Oliveira
Industrial batch processing is widely used in a number of areas of industrial production. In such processes, simultaneous and real-time measurements are taken from different process variables and large databases become thus available, enabling the precise monitoring of industrial operations. Data emerging from batch processes present special characteristics, like dynamic features, nonlinearity of data, non-gaussian data, etc; and there is therefore a growing interest in the development of customized multivariate control charts for their monitoring. We investigate an approach that uses control charts based on the STATIS method (from French: Structuration des Tableaux A Trois Indices de la Statistique), an exploratory technique for measuring similarities between data sets. Data are arranged in such a way that the monitoring along time is prioritized. The methodology easily allows a nonparametric on-line monitoring of complex batch processes in time, in situations where a large number of variables are present. The good performance of this approach is illustrated using simulated data.

E1760: **Monitoring occupational exposure data with joint control charts**
Presenter: **M Rosario Ramos**, FCiencias.ID, Portugal
Co-authors: Elisabete Carolino, Carla Viegas, Susana Viegas
Classical control systems using the Shewhart charts are often difficult to manage in real time and conditions of each environment. The use of univariate and independent control charts when measurements are relative to simultaneous observations, possibly correlated, distorts multivariate vector monitoring through type I error and the probability of being within the control limits will be not adjusted to the real value. This distortion of the control procedure increases with the number of variables to be controlled. Motivated by the need to promote a reliable monitoring system of the concentration of particles with adverse health effects, some alternatives are explored. A multivariate approach of control charts is used for the first time in the monitoring of the particulate matter in an animal feed production industry. Data are measurements on particles providing information about the concentration for particles of five diameters related with levels of severity of health effects. Location within the occupational environment is also known. Multivariate charts based on Hotellings T2 supplemented with graphics Shewhart and Bonferroni limits are applied as well as the generalization, to five variables, of the statistical decomposition of the Hotellings T2 for multiple comparisons. Some lines for further studies will be highlighted.

E1369: **Process monitoring for manufacturing attribute data using model-based approach**
Presenter: **Angelo Santanna**, Federal University of Bahia, Brazil
Solving problems in industry, even inside companies known as expert in their sector, is not just a question of applying the right technique. The control chart is a traditional tool for data monitoring processes and the model-based approach has been shown to be very effective in detecting disturbances in output variables when input variables are measurable. The idea of the model-based control chart is to integrate generalized linear model and control chart tools to monitor any changes in process data. In many situations, there are variables that are nonconforming data following a Binomial distribution, and the modeling and monitoring this type data suffers serious inaccuracies in control limits specification when the rate of nonconforming is small. We propose the monitoring of nonconforming data using a Beta distribution approximation. A case study is illustrated for the proposed method to compare the results against several model-based charts and a simulation study based on Markov chain is conducted to overcome such inaccuracies and performance for process monitoring.

CO723 Room Chancellor's Hall COINTEGRATION ANALYSIS AND STATE SPACE MODELS Chair: Martin Wagner

C1724: **A parameterization of MFI(1) and I(2) processes: Structure theory with an eye to hypothesis testing**
Presenter: **Dietmar Bauer**, University Bielefeld, Germany
Co-authors: Lukas Matuschek, Patrick de Matos Ribeiro, Martin Wagner

Important properties of a parameterization of state space models for unit root processes for the multiple frequency I(1) and I(2) cases are discussed. Given that formulating and estimating hypotheses on the cointegrating spaces is of particular importance, special emphasis is placed on describing the implementation of a variety of relevant hypotheses tests. A continuous parameterization necessarily partitions the set of all systems of a given minimal order into pieces, with the pieces described here by multi-indices (e.g., the unit root structure, Kronecker indices). The relation between the different pieces are discussed in detail, and it is shown, similar to the stationary case, that some pieces are at the boundary of other pieces. This structural information is of vital importance for parameter estimation and hypothesis testing in particular cointegrating ranks in, e.g., pseudo maximum likelihood estimation.

C1737: **Pseudo maximum likelihood analysis of multiple frequency I(1) processes: Parameter estimation**
Presenter: **Patrick de Matos Ribeiro**, Technical University Dortmund, Germany
Co-authors: Dietmar Bauer, Lukas Matuschek, Martin Wagner

The aim is to derive the asymptotic properties, consistency and the asymptotic distribution, of pseudo maximum likelihood parameter estimators for multiple frequency I(1) processes considered in the state space framework. With multiple frequency I(1) processes we denote processes with unit roots at arbitrary frequencies with integration orders all equal to one. As usual, the parameters corresponding to the nonstationary components are estimated super-consistently at rate T, whereas all other parameters are estimated at rate square root of T. The limiting distributions are mixtures of Brownian motions and normal distributions, respectively. Our simulation results indicate that for systems with unit roots at several frequencies pseudo maximum likelihood estimation leads to more precise estimates of the cointegrating spaces and to better predictions than subspace estimation of the state space systems or reduced rank regression of autoregressive approximations.

C1740: **Pseudo maximum likelihood analysis of multiple frequency I(1) processes: Inference on cointegrating ranks**
Presenter: **Lukas Matuschek**, Technical University Dortmund, Germany
Co-authors: Dietmar Bauer, Patrick de Matos Ribeiro, Martin Wagner

Based on an extension of the state space error correction model from the I(1) to the multiple frequency I(1) case, several test statistics are presented to determine the cointegrating ranks at the considered unit root frequencies. With multiple frequency I(1) processes we denote processes with unit roots at arbitrary frequencies with all integration orders equal to one. The considered test statistics differ whether additional parameter restrictions are imposed in a preliminary concentration step correcting for the other unit roots and stationary dynamics. However, all test statistics have the usual Brownian motion integral limiting null distributions. We complement our theoretical findings with a detailed simulation study. The most important finding is that for ARMA systems with roots close to the unit circle the state space model based test statistics outperform, in particular in small samples, previous tests calculated on autoregressive approximations with the lag lengths chosen by AIC.

C1633: **Long VAR approximation in I(2) context: Asymptotic theory and simulations**
Presenter: **Yuanyuan Li**, University of Bielefeld, Germany
Co-authors: Dietmar Bauer

The asymptotic theory for long VAR approximations is extended to I(2) processes. The analysis is mainly performed in the framework of a triangular representation admitting an infinite-order autoregressive representation subject to summability conditions on the autoregressive coefficients. The results, however, also have implications for more general data generating processes. Similar results as in I(1) cases are achieved including the consistency of the estimated coefficients as well as their asymptotic distributions for properly chosen lag length. Based on these results, tests for linear restrictions on the coefficients can be derived. The results are also the starting point for the derivation of rank tests and the asymptotic distributions of reduced rank estimators. Furthermore, a detailed simulation study examines the finite sample properties of rank testing procedures to specify the integer parameters (the two involved cointegration ranks) in the long VAR approximations for I(2) processes.

CO069 Room Court LATEST DEVELOPMENT OF DEPENDENCE MODELING Chair: Jean-David Fermanian

C0316: **Testing the simplifying assumption in high-dimensional vine copulas**
Presenter: **Fabian Spanhel**, LMU, Germany
Co-authors: Malte Kurz

Testing the simplifying assumption in high-dimensional vine copulas is a difficult task because tests must be based on estimated observations and amount to checking constraints on high-dimensional distributions. So far, corresponding tests have been limited to single conditional copulas with a low-dimensional set of conditioning variables. We propose a novel testing procedure that is computationally feasible for high-dimensional data sets and that exhibits a power that decreases only slightly with the dimension. By discretizing the support of the conditioning variables and incorporating a penalty in the test statistic, we mitigate the curse of dimensions by looking for the possibly strongest deviation from the simplifying assumption. The use of a decision tree renders the test computationally feasible for large dimensions. We derive the asymptotic distribution of the test and analyze its finite sample performance in an extensive simulation study. The utility of the test is demonstrated by its application to 10 data sets with up to 49 dimensions.

C0534: **About the estimation of the conditional Kendall's tau and Kendall's Regression**
Presenter: **Alexis Derumigny**, ENSAE-CREST, France
Co-authors: Jean-David Fermanian

The estimation of the conditional Kendall's tau is considered. The conditional Kendall's tau is a conditional dependence parameter between two variables conditionally to some observed covariates. We propose a nonparametric estimator using kernels. Under a pseudo-GLM specification, we also propose a parametric estimator for the (possibly sparse) vector of coefficients in the model. We study the theoretical properties of both estimators, and prove non-asymptotic bounds that holds with high probability.

C0551: **Simplified estimations of conditional copulas**
Presenter: **Jean-David Fermanian**, Ensae-Crest, France

Semiparametric conditional copula models suffer from the so-called curse of dimensionality. Indeed, conditional marginal distributions with a potentially large number of covariates have to be estimated with usual smoothing techniques. By assuming diverse single-index assumptions for such conditional distributions, we propose a simple way of reducing this curse of dimensionality. Therefore, with an underlying parametric conditional copula model under the simplifying assumption, some theoretical properties of estimated parameter are provided.

C0612: **Censored copula modeling for micro-level reserving in non life insurance**
Presenter: **Olivier Lopez**, Universita Pierre et Marie Curie Paris 6, France
In non life insurance, two main strategies exist to evaluate the amount of reserves required to cover the claims of the customers. Chain-Ladder approaches use aggregated information to forecast the amount to be paid, but they miss some important individual information one has on the claims. On the other hand, micro-level reserving aims to use each available information on the claim to produce individualized predictions of its final amount. Nevertheless, this second class of methods usually fails to achieve a reasonable performance for guarantees with potential large duration between the occurrence of the claim and the final payment (lifetime of the claim). We present a general method to address this issue, by modeling separately the hazard rate of the lifetime of the claim, and its amount, using a censored copula methodology to capture the dependence structure of these two variables.

CO081 Room G11 ECONOMETRIC MODELS FOR HIGH DIMENSIONAL SCORING Chair: Alessandra Amendola

C0428: **Contagion effects in small business failures: A spatial multilevel autoregressive model**
Presenter: **Raffaella Calabrese**, University of Edinburgh, United Kingdom
Co-authors: Robert Stine
The impact of nearby UK cities characteristics on small business defaults is studied. Credit scoring models would usually rely on city-level fixed effects to capture the economic conditions of the cities where SMEs are located. However, this method ignores the contagion effects given by network ties between neighboring cities. To include both contagion effects between and within cities, we propose a Bayesian multilevel model for binary data. We apply this model to data on SMEs located in the five biggest cities in the UK.

C0369: **Scoring models for roboadvisory platforms: A network approach**
Presenter: **Gloria Polinesi**, University of Ancona, Italy
Co-authors: Paolo Giudici
Due to technological advancement, roboadvisory platforms have allowed significant cost reduction in asset management. However, this improved allocation may come at the price of a biased risk estimation. To verify this, we empirically investigate allocation models employed by roboadvisory platforms. Our findings show that the platforms do not accurately assess risks and, therefore, the corresponding allocation models should be improved, incorporating further information, through clustering and network analysis.

C0825: **On the asymmetric impact of macro-variables on volatility**
Presenter: **Vincenzo Candila**, University of Salerno, Italy
Co-authors: Alessandra Amendola, Giampiero Gallo
The GARCH-MIDAS model is extended to take into account possible different impacts from positive and negative macroeconomic variations on financial market volatility. We evaluate the proposed specification by a Monte Carlo simulation which shows good estimation properties with the increase in the sample size. The empirical application is performed on the daily S&P 500 realized volatility dynamics with the monthly industrial production as an additional (signed) determinant. In the out-of-sample analysis, our asymmetric GARCH-MIDAS model statistically outperforms the competing specifications, represented by the GARCH(1,1), GJR-GARCH and GARCH-MIDAS models.

C1367: **On possible causal links between Twitter sentiment and banks financial ratios**
Presenter: **Paola Cerchiello**, University of Pavia, Italy
Co-authors: Juri Marcucci, Giancarlo Nicola, Giuseppe Bruno
The aim is to study the relationships between Twitter sentiment and various financial indicators (such as for example stock return or volume) of some of the major Italian banks. Moreover, we test the current technology for analyzing and evaluating the sentiment of short web-text messages written in Italian, such as those published on the Twitter micro-blogging platform. In fact, gauging the sentiment among financial investors is of paramount importance for both market participants and regulation authorities. Behavioural finance posits that stock market investors define their purchasing strategies considering arbitrage bounds and collective sentiments. Regulation and market authorities can address critical situations by collecting and analyzing the sentiment mood inferred from investors action on social media. The goal is twofold: on one hand, we provide an empirical method to evaluate the polarity of short text messages in Italian. Secondly, we establish a sound statistical framework to measure the causal links between sentiment extracted from Twitter and financial market variables even in presence of no stationarity and cointegration in the data. A quantitative evaluation of the impact of sentiment on financial indicators is relevant to increase the timely awareness of regulators with respect to potentially critical microeconomic conditions.

CO338 Room G5 EMPIRICAL MACROECONOMICS Chair: Christopher Otrok

C0512: **Structural interpretation of vector autoregressions with incomplete identification**
Presenter: **Christiane Baumeister**, University of Notre Dame, United States
Co-authors: James Hamilton
Traditional approaches to structural interpretation of vector autoregressions can be viewed as special cases of Bayesian inference arising from very strong prior beliefs about certain aspects of the model. These traditional methods can be generalized with a less restrictive Bayesian formulation that allows the researcher to summarize uncertainty coming not just from the data but also uncertainty about the model itself. We use this approach to revisit the role of shocks to oil supply and demand and conclude that oil price increases that result from supply shocks lead to a reduction in economic activity after a significant lag, whereas price increases that result from increases in oil consumption demand do not have a significant effect on economic activity.

C0515: **An empirical investigation of direct and iterated multistep approaches to producing conditional forecasts**
Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States
When constructing an unconditional point forecast, both direct and iterated multistep (DMS and IMS) approaches are common. However, in the context of producing conditional forecasts, IMS approaches based on vector autoregressions (VAR) are far more common than simpler DMS, horizon-specific autoregressive-distributed lag (ARDL) models. This is despite the fact that there are theoretical reasons to believe that DMS methods are more robust to misspecification than are IMS methods. In the context of unconditional forecasts, the empirical relevance of these theories has been investigated. We extend that work to conditional forecasts. We do so based on linear bivariate and trivariate VARs/ARDLs estimated using a large dataset of macroeconomic time series.

C0559: **Mending the broken link: Heterogeneous bank lending and monetary policy pass-through**
Presenter: **Matteo Ciccarelli**, European Central Bank, Germany
The purpose is to analyze the pass-through of monetary policy measures to lending rates to firms and households in the euro area using a unique bank-level dataset. Bank balance sheet characteristics such as the capital ratio and the exposure to sovereign debt are responsible for the heterogeneity of pass-through of conventional monetary policy changes. The location of a bank is instead irrelevant. Non-standard measures normalized

the capacity of banks to grant loans resulting in a significant compression in lending rates. Banks with a high level of non-performing loans and a low capital ratio were the most responsive to the measures. Finally, we quantify the effects of non-standard policies on the real economic activity using a standard macroeconomic model and find that in absence of these measures both inflation and output would have been significantly lower.

C0541: **A comprehensive view of income inequality**
Presenter: **Christopher Otrok**, University of Missouri and FRB St Louis, United States
A new dataset is built that measures income percentiles by demographic group. Our groups are based on gender, education and race. We find that there are large differences in changes in income inequality across demographic groups. Further, the role that demographic factors play depends on the percentile of the income distribution. We then develop a dynamic factor model that captures both demographic-specific factors as well as income-quantile specific factors. We then embed this factor model in a VAR to identify the structural sources of inequality by demographic/income group.

CO102 Room Gordon COMPUTATIONAL ECONOMETRICS AND MODELLING II Chair: Gareth Peters

C0553: **Modelling the limit order book using marked Hawkes self-exciting point processes.**
Presenter: **Kylie-Anne Richards**, University of New South Wales, Australia
Increased activity and temporal clustering in the limit order book (LOB) can be characterized by an increase in intensity of events. Understanding and forecasting fluctuations in the intensity is informative to high frequency financial applications. The Hawkes self-exciting point process can be used to successfully model the dynamics of the intensity function by allowing for irregularly spaced time sequences, a multivariate framework, multiple dependent marks and the ability to capture the impact of marks on intensity. A critical first step to successfully apply these models to the LOB is suitably defining events in terms of the number of limit order book levels and types of orders. Based on extensive data analysis, recommendations are made. Likewise, selection of marks that impact the intensity function is challenging and the literature provides little guidance. Based on a review of the LOB literature potential marks are identified and screened using a novel mark detection method based on the likelihood score statistic. Comparisons of exponential and power-law decay functions are presented. Fitting marks with a likelihood based method presents substantial identifiability issues which are investigated via simulation for a variety of model formulations. Application is made to futures data with various underlying asset classes.

C0722: **Efficiently estimating discrete and continuous time GARCH models with irregularly spaced observations**
Presenter: **William Dunsmuir**, The University of New South Wales, Australia
There are essentially two continuous time limits of GARCH(1,1) processes as the time between observations shrinks to zero. The first, the bivariate diffusion limiting process does not allow jumps in the continuous time limit process. The COGARCH process defined in terms of a Levy process can also be obtained as the continuous time limit of the discrete time GARCH(1,1) process. The COGARCH process has a single source of driving noise and allows jumps. Because the number, time location and size of jumps cannot be observed directly using equally or irregularly spaced observations on the continuous time process the likelihood for the COGARCH model is intractable and requires careful computational implementation. Sequential Monte Carlo (SMC) with a continuous resampling method to estimate the likelihood function and ensure it is continuous in the parameters is used. We show that the SMC based method outperforms the quasi-maximum likelihood methods previously proposed in the literature in terms of bias and standard errors of estimation. Application to high frequency financial returns data will be presented. The SMC approach can also be used to estimate the parameters of traditional discrete time GARCH models and variants such as the Markov regime switching GARCH model when they are irregularly observed. Illustrations for financial time series as well as high frequency wind measurements will also be presented.

C1313: **An enhance empirical mode decomposition**
Presenter: **Marta Campi**, University College London, United Kingdom
Co-authors: Gareth Peters, Nourddine Azzaoui, Tomoko Matsui
A new technique, called Enhanced Empirical Mode Decomposition (EEMD), will be presented. Firstly, we introduce which are the classical a-priori decomposition basis techniques within the literature such as the Fourier Transform, the Wavelet Transform. Afterwards, the classical Empirical Mode Decomposition is presented. We will then underline pros and cons of it by paying particular attention to its drawbacks and so highlighting the motivation behind my new technique. Afterwards, a finance application comparing the EMD and the EEMD is provided in order to present its main advantages.

C1317: **Framework for detecting statistical causality in warped Gaussian processes**
Presenter: **Anna Zaremba**, University College London, United Kingdom
Co-authors: Gareth Peters
A new framework is proposed for detecting statistical causality. The challenge in identification and detection of causality in multivariate, non-linear, non-Guassian time series that exhibit non-linear dependence is not trivial. Consequently, models that would allow for structural properties like: non-stationarity, heteroscedasticity, tail dependence, long memory, are rarely studied, especially in the context of statistical causality testing and inference. Gaussian processes are a flexible class of models that can be used to semi-parametrically model time series observations and can incorporate different structural properties. We consider deformations of Gaussian processes, sometimes referred to as warped Gaussian processes, that are specifically designed in this context to capture multivariate dependence and concordance relationships that, when not included in the time series data model, may obfuscate the ability to detect linear and non-linear causal relationships between each marginal time series in the drift or volatility. Testing is performed using the generalised likelihood ratio test (GLRT), and under the assumptions of nested models, where the null hypothesis is the one of lack of causality. We describe ways of constructing such tests to be applicable for a wide range of data structures.

CO300 Room Jessel NEW APPROACHES TO MACROECONOMIC ANALYSIS AND FORECASTS Chair: Deborah Gefang

C1141: **Investigating strong-form purchasing power parity for EUR/PLN within STVECM framework**
Presenter: **Adrian Burda**, Uniwersytet Ekonomiczny w Krakowie, Poland
Existence of strong-form purchasing power (PPP), both for developed and emerging economies remains a widely investigated topics in international finance and macroeconomics. In case of Polish zloty, before the global financial crisis the existence strong-form PPP had been rejected in most studies, due to its appreciation against to main currencies. However, after 2008 more and more arguments supporting this hypothesis occured. The purpose is to verify strong-form purchasing power parity (PPP) of EUR/PLN within different variants of exponential smooth transition vector error correction model (ESTVECM) as well as first and second order logistic smooth transition vector error correction framework (LSTVECM and 2LSTVECM respectively). Significance of respective exponential and logistic smooth transition functions is compared with linear error-correction mechanism mechanism present in VECM. The competitive models for recursive samples are compared by likelihood ratio test, information criteria, out of sample forecast accuracy measures and by directional accuracy test.

C1186: **Synchronization of cycles in a data-rich environment**
Presenter: **Cem Cakmakli**, Koc University, Turkey
Co-authors: Richard Paap
A general framework is proposed where the synchronization of the cycles embedded in multiple leading and coincident variables are modeled jointly. We use dynamic factor structure together with Markov mixture models to estimate the coincident and leading economic factors. The novel feature of the model is that we can estimate the coincident and leading economic factor together with their lead/lag relationship in a time varying manner. We show that lead time of the leading economic factor changes drastically over time. While the lead time increases during recession periods, it drops considerably at the onset of expansions. Finally, we provide evidence on superior predictive capability of our model over existing popular procedures in predicting key US macroeconomic variables.

C1272: **Bayesian Lasso for large vector autoregression models**
Presenter: **Deborah Gefang**, University of Leicester, United Kingdom
Recent years have witnessed a growing interest in using Bayesian vector autoregression model (BVAR) that consists of many variables to gain deeper insight from large macroeconomic and financial data sets. Most of the available Bayesian techniques for large BVAR models, however, are subject to two types of constraints. The first type arises from the use of subjective priors which might not always be in line with the data generating process. The second type of constraint is associated with the computing power of the available computers, including the high performance computers. A Bayesian LASSO VAR method is developed to tackle these problems. Using FRED-MD monthly data for Macroeconomic Research provided by McCracken, we provide a comprehensive comparison between the forecasting performance of our method and that of other popular estimation techniques.

C0691: **Optimal window selection for forecasting in the presence of recent structural breaks**
Presenter: **Yongli Wang**, University of Leicester, United Kingdom
Two feasible algorithms are proposed to select the optimal window size for forecasting in rolling regression. The proposed methods are developed based on the existing methodology, keeping the asymptotic validity and allowing for the lagged dependent variable in regression and multi-step ahead forecasting. The Monte-Carlo experiments show that the proposed bootstrap method outperforms the original algorithm in the literature in almost all cases. It is also shown that the forecasts from the proposed methods are superior to those from other existing methods in some cases, and close to the best forecasts in other cases. However, when the break occurs far before the time of making forecasts and the break size is significant, using only post-break data is almost always the best strategy.

CO543 Room Montague RECENT DEVELOPMENTS IN ECONOMETRICS OF ASSET PRICING Chair: Guillaume Roussellet

C1422: **European spreads at the interest rate lower bound**
Presenter: **Laura Coroneo**, University of York, United Kingdom
Co-authors: Sergio Pastorello
The effect of the interest rate lower bound on long-term sovereign bond spreads in the Euro area is analyzed. We specify a joint shadow rate model for the risk-free, the German and the Italian yield curves. Results indicate that in 2012 the 10-year Italian sovereign spread with respect to Germany was constrained by the lower bound on interest rates by as much as 0.77%. We show that this nonlinearity implies that, at the interest rate lower bound, long-term sovereign bond spreads are asymmetrically distributed, affected by changes in the lower bound and less informative about the sovereign risk. Shadow spreads, however, still provide reliable information.

C1820: **Crash risk in individual stocks**
Presenter: **Paola Pederzoli**, University of Geneva Swiss Finance Institute, Switzerland
Crash risk in individual stocks is studied using a large cross-section of options and find evidence of a significant positive skewness risk premium since the financial crisis of 2008/2009. Using a novel trading strategy in form of a skewness swap, we document that the average monthly return of the skewness swap is 50% with an annualised Sharpe ratio of 1.16 and that these returns remain robust when taking into account transaction costs. Moreover, we find that only the idiosyncratic risk premium increases after the crisis while the coskewness risk premium remains at pre-crisis levels. Finally, we theoretically and empirically link the skewness risk premium to short-sell constraints in a framework where investors face asymmetric information.

C1787: **Efficient parameter estimation for multivariate jump-diffusions**
Presenter: **Gustavo Schwenkler**, Boston University, United States
Co-authors: Francois Guay
Unbiased estimators of the transition density and posterior filters of a multivariate jump-diffusion process are developed. The drift, volatility, jump intensity, and jump magnitude are allowed to be state-dependent and non-affine. It is not necessary to diagonalize the volatility matrix. Our approach facilitates the parametric estimation of multivariate jump-diffusion models based on discretely observed data. Our parameter estimators have the same asymptotic behavior as maximum likelihood estimators under mild conditions. Our methodology is found to be highly accurate and computationally efficient for the estimation of consumption growth dynamics.

C1802: **Affine modelling of credit risk, pricing of credit events and contagion**
Presenter: **Guillaume Roussellet**, McGill University, Canada
Co-authors: Alain Monfort, Jean-Paul Renne, Fulvio Pegoraro
A new discrete-time affine pricing model for defaultable securities is proposed breaking down the most restrictive assumptions made in existing frameworks. Specifically, our model simultaneously allows for the presence of systemic entities by departing from the no-jump condition on the factors conditional distribution, contagion effects, the pricing of credit events and (iv)the presence of stochastic recovery rates. Our affine framework delivers explicit pricing formulas for default-sensitive securities like bonds and credit default swaps. A first application shows how this framework can be exploited to estimate sovereign credit risk premiums in an equilibrium model. In a second application, we jointly model term structures of sovereign CDS denominated in different currencies and extract market-implied probabilities of depreciations at default. A third application illustrates the ability of the model to replicate the behavior of banks CDS spreads that was observed in the aftermath of the Lehman Brothers bankruptcy.

CO417 Room Senate MODEL SELECTION FACING SHIFTS Chair: David Hendry

C0427: **Robust model selection: A review**
Presenter: **Jennifer Castle**, Oxford University, United Kingdom
Co-authors: David Hendry
Complete and correct specifications of models for observational data never exist, so model selection is unavoidable. The target of selection needs to be the process generating the data for the variables under analysis, while embedding the objective of the study, often a theory-based formulation. This requires starting from a sufficiently general initial specification that comprises all candidate variables, their lags in time-series data, and functional forms, allowing for possible outliers and shifts, seeking parsimonious final representations that retain the relevant information, are well specified, encompass alternative models, and evaluate the validity of the objective. Intrinsically, we seek robustness against many potential problems jointly: outliers, shifts, omitted variables, incorrect distributional shape, non-stationarity, mis-specified dynamics, and non-linearity, as well as inappropriate exogeneity assumptions. Our approach inevitably leads to more variables than observations, tackled by iteratively switching between contracting and expanding multi-path searches programmed in Autometrics. The steps involved are explained, specifically addressing indicator saturation to discriminate between outliers and large observations arising from non-linear responses. The analysis is illustrated using artificial data to evaluate outliers versus non-linearity and by a model of engine knock in which there is incorrectly recorded data, identifying shifts in relations.

C0377: **On the ability to adapt to changes: An assessment of hurricane damage mitigation efforts using forecast uncertainty**
Presenter: **Andrew Martinez**, University of Oxford, United Kingdom
Humanity's ability to adapt to extreme weather events is increasingly relevant due to climate volatility. Given the uncertainty associated with climate outcomes, it is important to understand the effectiveness of adaptation across various horizons. However, there is little analysis of the effectiveness of short-term adaptation efforts. We propose a novel approach of using errors from hurricane forecasts to evaluate hurricane damage mitigation efforts in the short-term. We construct a general model of hurricane damages for all hurricanes to strike the continental United States in the past 60 years. While we allow for both existing explanations of hurricane impacts as well as additional determinants, we find that a small subset explains most of the variation in hurricane damages. This result is broadly robust to outliers and shifts.

C0437: **Asymptotic theory of M-estimators for linear regression in time series**
Presenter: **Bent Nielsen**, University of Oxford, United Kingdom
Co-authors: Soren Johansen
An asymptotic theory is provided for a class of regression M-estimators. The objective function must be continuous, but it can be non-convex and non-differentiable. The regression equation has innovations that can have a continuous or discrete distribution. The regressors must satisfy an assumption on the frequency of small regressors. This is met by a variety of deterministic or stochastic regressors, including stationary and random walk regressors. In a previous paper we have shown that condition along with some mild conditions on the criterion function is sufficient to ensure boundedness, or tightness, of their non-standardized distributions.

C0380: **First-in, first-out: Modelling the UK's CO2 emissions, 1860–2016**
Presenter: **David Hendry**, Oxford, United Kingdom
The United Kingdom was the first country into the Industrial Revolution in the mid-18th Century, building on earlier revolutions in scientific, technological and medical knowledge. 250 years later, real income levels in the UK are about 7-10 fold higher per capita, even greater in some other countries, many killer diseases have been tamed, and longevity has approximately doubled. However, such beneficial developments have led to an explosion in anthropogenic emissions of carbon dioxide and other greenhouse gases. The UK now seems to be one of the first countries out, as its CO2 emissions are back to 1890's levels. We develop an econometric model of that process over the last 150 years, and confirm the key role of reduced coal use, and also of the capital stock, which embodies the vintage of technology at its construction. Many major shifts and outliers require to be handled to develop a viable model.

CO087 Room Woburn MULTIVARIATE MODELLING OF ECONOMIC AND FINANCIAL TIME SERIES Chair: Alain Hecq

C0248: **The empirical effects of anticipated and unanticipated government spending shocks in the United States**
Presenter: **Thomas van Gemert**, Maastricht University, Netherlands
Co-authors: Lenard Lieb
The dynamic effects of anticipated and unanticipated government spending shocks in the United States are investigated. To this aim, we link narrative records on defense spending news to the legislative decision process and use the lag between the announcement date in the narrative record and actual legal implementation to differentiate between unanticipated and anticipated shocks. We find empirical evidence that unanticipated spending shocks result in higher multiplier effects than anticipated ones. Private consumption and investment are crowded-out by anticipated shocks, but not by unanticipated shocks.

C0310: **On cointegration for processes integrated at different frequencies**
Presenter: **Tomas del Barrio Castro**, University of the Balearic Islands, Spain
Co-authors: Gianluca Cubadda, Denise Osborn
The possibility of having cointegration relationships between processes integrated at different frequencies is explored. We found that it is possible to establish this kind of cointegration and that the only possible cointegration relationships between processes integrated at different frequencies is full periodic polynomial cointegration. We explore the connection of this kind of cointegration with the demodulator operator and finally propose a simple way to test for the presence of cointegration between processes integrated at different frequencies based on the use of the demodulator operator.

C1394: **A vector index-augmented heterogeneous autoregressive model for forecasting realized covariance matrices**
Presenter: **Alain Hecq**, Maastricht University, Netherlands
A multivariate model for the elements of realized covariance matrices is proposed where each equation follows a heterogeneous autoregressive model that is augmented with a common index structure. Our modelling can accommodate both idiosyncratic and common dynamics of realized covariances. We offer a switching algorithm to maximise the Gaussian likelihood of our model. Since the maximum likelihood estimator may perform poorly when the dimension of the covariance matrix becomes large, we also propose some algorithms for regularized estimation. The proposed approach is evaluated through both simulations and empirical examples.

C1838: **Incentive-driven inattention**
Presenter: **Joao Victor Issler**, Getulio Vargas Foundation, Brazil
Co-authors: Wagner Gaglianone, Raffaella Giacomini, Vasiliki Skreta
A dynamic rational inattention model is built where agents have a budget of attention or cognitive resources to devote to updating expectations.

Both the amount of attention and the expectations' accuracy are endogenous and linked to the cost and benefit of updating, which can vary across agents and over time. The model captures novel stylized facts from a panel of professional forecasters, the Central Bank of Brazil's Focus Survey, that is the unique among surveys in that agents choose when to update and there is a recurring contest that ranks agents based on their accuracy. We find that the incentives linked to the contest are the primary drivers of updates and accuracy gains. We structurally estimate the model to uncover the deep parameters characterizing the cost-benefit distribution across agents and how this changes around the contest. The model fits the data well and the estimates allow us to perform counter factual exercises to understand the value of the contest and to investigate alternative survey designs.

CO346 Room SH349 NON STANDARD INFERENCE PROBLEMS IN ECONOMETRICS Chair: Lionel Truquet

C0958: **A simple and practical approach towards testing global restrictions on general functions**
Presenter: **Valentin Patilea**, CREST-Ensai, France
Co-authors: Jeffrey Racine
A simple bootstrap procedure is proposed for inference on vectors or functions in a general context that involves estimation only under the alternative, while constraints are imposed via choice of a suitable transformation of the unconstrained estimate. The procedure is quite general and applies directly to functions or derivatives defined by separable and non-separable regression models. It can be used with parametric, semi and nonparametric estimators without modification. Potential applications include, but are not limited to, inequality inference on mean or quantile regression models where the bounds depend on the model's covariates, checking monotonicity, convexity, symmetry, homogeneity, for multivariate functions.

C1219: **Robust inference in differentiated products demand models**
Presenter: **Pujee Tuvaandorj**, CREST-ENSAI, France
Co-authors: Stephane Auray, Nicolas Lepage-Saucier
Robust inference in random coefficient logit models for differentiated products demand is studied. The model is subject to two different irregularities that may lead to a failure of standard inference: (i) the variance of the random coefficients is often close to zero (implying little variation in tastes), which leads to the boundary parameter problem; and (ii) the strength of the available instruments is often put in doubt, which may cause weak identification. We construct test statistics that simultaneously overcome both types of irregularities. The test statistics are asymptotically pivotal i.e., their asymptotic distribution does not depend on unknown parameters irrespective of the identification strength of the instruments and the degree of heterogeneity. We evaluate the performance of the inference procedures through simulations and present an application to the U.S. automobile market.

C1036: **Response error and transformation models**
Presenter: **Stephan Martin**, Humboldt University Berlin, Germany
Co-authors: Christoph Breunig
Nonclassical measurement error of the dependent variable is considered. We show that this specification leads to a transformation model. Identification arguments vary for continuous and mixed observed dependent variables. Nonparametric identification of the regression function is achieved. We also provide a novel estimation strategy for the transformation model.

C1402: **Inference after model averaging in linear regression models**
Presenter: **Chu-An Liu**, Academia Sinica, Taiwan
Co-authors: Xinyu Zhang
The problem of inference for nested least squares averaging estimators is considered. We study the asymptotic behavior of the Mallows model averaging estimator and the jackknife model averaging estimator under the standard asymptotics with fixed parameters setup. We find that both MMA and JMA estimators asymptotically assign zero weight to the under-fitted models, and MMA and JMA weights of just-fitted and over-fitted models are asymptotically random. Building on the asymptotic behavior of model weights, we derive the asymptotic distributions of MMA and JMA estimators and propose a simulation-based confidence interval for the least squares averaging estimator. Monte Carlo simulations show that the coverage probabilities of proposed confidence intervals achieve the nominal level.

CG247 Room G4 CONTRIBUTIONS IN VOLATILITY MODELING Chair: Francesco Audrino

C1419: **Asymmetric and symmetric volatility models for exchange rates in India**
Presenter: **Anand Shah**, Tata Consultancy Services (TCS), India, India
The aim is to model the monthly and the daily US, Euro Zone, UK and Australian exchange rates in India using the symmetric (sGARCH) and the asymmetric (GJR-GARCH and EGARCH) volatility models with the normal, the student t and the skewed student t error distributions. We also investigate the effect of the net US dollars (USD) purchase/ sale by the central bank, the net foreign institutional investor (FII) inflows and the one month forward spot differential on the monthly US exchange rate. Furthermore, we also test the presence of the calendar effect such as the monthly effect and the day-of-week effect on these exchange rates. The calendar effects in the mean model are pronounced and Fridays on an average witness an appreciation of the Indian rupee against all the currencies we tested. Month of August has significant impact on the US exchange rate and June on the other three exchange rates. The leverage effect is not pronounced in all the exchange rates. The concern that capital inflows, especially the easily repatriable ones, could appreciate the Indian rupee seems to be correct but the net purchase/ sale of USD by the RBI does not seem to be abating the impact.

C1669: **Uncertainty and economic activity: Identification through cross-country correlations**
Presenter: **Ambrogio Cesa-Bianchi**, Bank of England, United Kingdom
Co-authors: Hashem Pesaran, Alessandro Rebucci
A multi-country approach is proposed for the analysis of the interaction between uncertainty and economic activity both within and across economies. Two global factors, labeled real and financial, are considered and, assuming different patterns of cross-country, correlations of country-specific innovations to output growth and stock market volatility are identified. It is assumed that only the real factor is required to model cross-country correlations of growth innovations, but both factors are needed to model cross-country correlations of volatility innovations. These assumptions are shown to be in accordance with stylized facts of the data and the estimated innovations. The identified factors are then used in a factor-augmented VAR to quantify the impact of common and country-specific shocks. It is found that most of the unconditional correlation between volatility and growth can be accounted for by the two common factors. While unconditionally volatility behaves countercyclically, once we condition on the factors the growth-volatility correlation becomes essentially negligible. The share of volatility variance explained by the real factor and by country-specific growth shocks is shown to be relatively small. Similarly, shocks to the financial factor explain only a small fraction of the country-specific growth variance. Finally, country-specific shocks are important for domestic volatility and growth, but have limited spillover effects to other countries.

C1698: **Asymptotic theory of the QMLE of the EGARCH-type models**
Presenter: **Xiaoyu Li**, Capital University of Economics and Business, China
Co-authors: James Davidson
The asymptotic properties of the quasi-maximum-likelihood estimator (QMLE) for the EGARCH-type models, including the EGARCH, HYEGARCH and FIEGARCH(DL) models, are investigated. We first review the literature on the invertibility and asymptotic properties of the EGARCH(1,1) processes and establish the consistency of the QMLE in the HYEGARCH models under mild condition. We also provide an investigation into the asymptotic normality of the QMLE of the HYEGARCH and FIEGARCH(DL) processes. Finally, we demonstrate the finite sample properties of the QMLE for the HY/FIEGARCH(DL)(0,d,0) processes through a Monte Carlo simulation.

C1675: **Volatility forecasting with empirical similarity: Japanese stock market case**
Presenter: **Takayuki Morimoto**, Kwansei Gakuin University, Japan
Co-authors: Yoshinori Kawasaki
The forecasting ability of various volatility models is compared through within-sample and out-of-sample forecasting simulations. The considered models are heterogeneous auto-regression models (HAR), a 1/3 model where the weight coefficients are all set to 1/3 in the HAR model (ES0), and an HAR model where the weight coefficients are determined by their empirical similarity. We also test AR(1), ARCH/GARCH and their variants, and models incorporating the realized quarticity (RQ), which are referred to as ARQ, HARQ, and ESQ. For stock data, we picked six index series stocks that are listed on the Tokyo Stock Exchange as well as 24 individual stock series. All these stocks had enough liquidity in the market from April 1, 1999, to December 30, 2013, for our investigation. Minute-by-minute data were created based on high-frequency data. Forecasting evaluation depends on what kind of evaluation function we employ. We make use of Patton's error function. By changing the length of estimation period and the forecasting period and the parameter of Patton's error function, we attempt 27,000 forecasting simulations. We find that ESQ and HARQ are almost comparative in within-sample forecasting, whereas ES0 differs in out-of-sample forecasting experiments. We also tried a model comparison based on a previous pair-wise testing procedure. We found similar results, but the details are different between the index series and the individual stock series.

Monday 18.12.2017 08:40 - 10:20 Parallel Session L – CFE-CMStatistics

EO498 Room CLO B01 RECENT ADVANCES OF BAYESIAN METHODS FOR COMPLEX DATA ANALYSIS Chair: Weining Shen

E0533: **On distributed Bayesian computation**
Presenter: **Botond Szabo**, Leiden University, Netherlands
Co-authors: Harry Zanten
First, the theoretical properties of various Bayesian distributed methods are investigated on the benchmark signal in Gaussian white noise model. Then we consider the limitations and guarantees of distributed methods in general under communication constraints on the same benchmark nonparametric model.

E1099: **Validating approximate Bayesian computation on posterior convergence**
Presenter: **Wentao Li**, Newcastle University, United Kingdom
Co-authors: Paul Fearnhead
Many statistical applications involve models for which it is difficult to evaluate the likelihood, but relatively easy to sample from. Approximate Bayesian computation is a likelihood-free method for implementing Bayesian inference in such cases. We present a number of surprisingly strong asymptotic results for the regression-adjusted version of approximate Bayesian Computation. We show that for an appropriate choice of the bandwidth in approximate Bayesian computation, using regression-adjustment will lead to a posterior that, asymptotically, correctly quantifies uncertainty. Furthermore, for such a choice of bandwidth we can implement an importance sampling algorithm to sample from the posterior whose acceptance probability tends to 1 as we increase the data sample size. This compares favourably to results for standard approximate Bayesian computation, where the only way to obtain its posterior that correctly quantifies uncertainty is to choose a much smaller bandwidth, for which the acceptance probability tends to 0 and hence for which Monte Carlo error will dominate.

E1177: **On the Pitman-Yor process with spike and slab prior specification**
Presenter: **Bernardo Nipoti**, Trinity College Dublin, Ireland
Co-authors: Igor Pruenster, Antonio Lijoi, Antonio Canale
For the most popular discrete nonparametric models, beyond the Dirichlet process, the prior guess at the shape of the data generating distribution, also known as base measure, is assumed to be diffuse. Such a specification greatly simplifies the derivation of analytical results allowing for a straightforward implementation of Bayesian nonparametric inferential procedures. However, in several applied problems the available prior information leads naturally to incorporate an atom into the base measure and one is essentially left with the Dirichlet process as the only tractable choice for the prior. We fill this gap by considering the Pitman-Yor process featuring an atom in its base measure. We derive computable expressions for the distribution of the induced random partitions and for the predictive distributions. These findings allow us to devise an effective generalized Polya urn Gibbs sampler. Applications to density estimation, clustering and curve estimation, with both simulated and real data, serve as an illustration of our results and allow comparisons with existing methodology. In particular, we tackle a functional data analysis problem concerning basal body temperature curves.

E1013: **Frequentist coverage and sup-norm convergence rate in Gaussian process regression**
Presenter: **Yun Yang**, Florida State University, United States
Co-authors: Debdeep Pati, Anirban Bhattacharya
Gaussian process (GP) regression is a powerful interpolation technique due to its flexibility in capturing non-linearity. We provide a general framework for understanding the frequentist coverage of point-wise and simultaneous Bayesian credible sets in GP regression. As an intermediate result, we develop a Bernstein von-Mises type result under supremum norm in random design GP regression. Identifying both the mean and covariance function of the posterior distribution of the Gaussian process as regularized M-estimators, we show that the sampling distribution of the posterior mean function and the centered posterior distribution can be respectively approximated by two population level GPs. By developing a comparison inequality between two GPs, we provide exact characterization of frequentist coverage probabilities of Bayesian point-wise credible intervals and simultaneous credible bands of the regression function. Our results show that inference based on GP regression tends to be conservative; when the prior is under-smoothed, the resulting credible intervals and bands have minimax-optimal sizes, with their frequentist coverage converging to a non-degenerate value between their nominal level and one. As a byproduct of our theory, we show that the GP regression also yields minimax-optimal posterior contraction rate relative to the supremum norm, which provides a positive evidence to the long standing problem on optimal supremum norm contraction rate in GP regression.

EO196 Room MAL B18 NEW DEVELOPMENT IN CAUSAL INFERENCE Chair: Xavier De Luna

E0965: **Tests for the unconfoundedness assumption using quasi-instruments**
Presenter: **Emma Persson**, Umea University, Sweden
Co-authors: Xavier de Luna, Per Johansson
In observational studies, the identification of an average causal effect of a treatment on an outcome of interest commonly relies on the unconfoundedness assumption. An alternative approach to identification is to use an instrument. When an instrument is available, it may also be used to test for the unconfoundedness assumption even in situations where the instrument does not fulfill all necessary conditions to yield nonparametric identification of the average causal effect. We propose tests for the unconfoundedness assumption using such quasi-instrumental variables, both using matching and parametric models. Our approach allows for discrete as well as continuous instruments, and is applicable to situations where the targeted parameter is the average treatments effect on the treated. We perform a simulation study to evaluate the finite sample performance and compare power with Durbin-Wu-Hausman tests. Finally we apply the results to a case study where the interest lies in evaluating the effect of cognitive behavioral therapy on three outcomes; outpatient care, sickness leave and medical drug prescriptions.

E1264: **Multiple questions for multiple mediators**
Presenter: **Bianca Lucia De Stavola**, University College London GOS Institute of Child Health, United Kingdom
Co-authors: Rhian Daniel, Nadia Micali, George Ploubidis
Investigating the mechanisms that may explain the causal links between an exposure and a temporally distal outcome often involves multiple interdependent mediators. Until recently, dealing with multiple mediators was restricted to settings where mediators relate to exposure and outcome only linearly. Extensions proposed in the causal inference literature to allow for interactions and non-linearities in the presence of multiple mediators initially focussed on natural direct and indirect effects. These however are not all identifiable, with the rest requiring stringent, and often unrealistic, assumptions. More recent developments have focussed interventional (or randomised interventional analogue) direct and indirect effects to deal with these issues. They can be identified under less restrictive assumptions, with generalizations dealing with time-varying exposures, mediators and confounders also possible. The mediation questions that can be addressed when estimating interventional effects differ from those asked by natural effects in subtle ways. We will review them, discuss their differences in emphasis, assumptions, and interpretation, and propose ways

of exploiting these differences to assess the robustness of conclusions. We will use an epidemiological investigation of the mechanisms linking maternal pre-pregnancy weight status and offspring eating disorders behaviour to illustrate these points.

E0902: **Bayesian nonparametric generative models for causal inference with missing at random covariates**
Presenter: **Michael Daniels**, University of Texas at Austin, United States
Co-authors: Jason Roy
A general Bayesian nonparametric (BNP) approach is proposed for causal inference in the point treatment setting. The joint distribution of the observed data (outcome, treatment, and confounders) is modeled using an enriched Dirichlet process. The combination of the observed data model and causal assumptions allows us to identify any type of causal effect - differences, ratios, or quantile effects, either marginally or for subpopulations of interest. The proposed BNP model is well-suited for causal inference problems, as it does not require parametric assumptions about the distribution of confounders and naturally leads to a computationally efficient Gibbs sampling algorithm for both large n and p. By flexibly modeling the joint distribution, we are also able to impute (via data augmentation) values for missing covariates within the algorithm under an assumption of ignorable missingness, obviating the need to create separate imputed data sets. This approach for imputing the missing covariates has the additional advantage of guaranteeing congeniality between the imputation model and the analysis model, and because we use a BNP approach, parametric models are avoided for imputation. The performance of the method is assessed using simulation studies. The method is applied to data from a cohort study of human immunodeficiency virus/hepatitis C virus co-infected patients.

E0720: **New matching methods for causal inference using integer programming**
Presenter: **Jose Zubizarreta**, Harvard University, United States
Co-authors: Magdalena Bennett, David Hirshberg, Juan Pablo Vielma
In observational studies of causal effects, matching methods are often used to approximate the ideal study that would be conducted if it were possible to do it by controlled experimentation. We will discuss new matching methods based on integer programming that allow the investigator to overcome three limitations of standard matching approaches by: (i) directly obtaining flexible forms of covariate balance; (ii) producing self-weighting matched samples that are representative by design; and (iii) handling multiple treatment doses without resorting to a generalization of the propensity score. (iv) Unlike standard matching approaches, with these new matching methods typical estimators are root-n consistent under the usual conditions. We will illustrate the performance of these methods in real and simulated data sets.

EO413 Room MAL B20 SPATIAL EXTREMES Chair: Marco Oesting

E0814: **Extremal (in)dependence structures of copulas with multiplicative constructions**
Presenter: **Sebastian Engelke**, Ecole Polytechnique Federale de Lausanne, Switzerland
Co-authors: Thomas Opitz, Jenny Wadsworth
A main construction principle for bivariate copulas with desirable tail properties uses the multiplicative representation $R(W_1,W_2)$, where R is a univariate scaling variable, and $W=(W_1,W_2)$ is a bivariate random vector. Numerous models in extreme value statistics are particular cases of this construction, and, depending on the distributions of R and W, they can result in either asymptotic dependence or asymptotic independence. We systematically characterize the extremal dependence structures arising from such multiplicative constructions. It turns out to be crucial how the tail decay rate in R impacts the tail dependence of (W_1,W_2). The results allow us to recover the extremal properties of existing models in a unified way, and, on the other hand, they can be used to construct new statistical models with flexible tail (in)dependence structures. The theory can also be applied to understand tail properties of spatial models.

E1150: **A central limit theorem for functions of stationary max-stable processes on $\mathbb{R}^d$**
Presenter: **Erwan Koch**, ISFA and CREST, France
Co-authors: Clement Dombry, Christian Y. Robert
Max-stable processes are very appropriate for the statistical modeling of spatial extremes. We propose a central limit theorem for functions of stationary max-stable processes on R^d. As an application of this result, we show the asymptotic normality of three simple estimators of the pair extremal coefficient.

E1132: **Multivariate regular variations and the Husler-Reiss Pareto model**
Presenter: **Clement Dombry**, Universite de Franche Comte, France
Co-authors: Olivier Zhen Wai Ho
A simple procedure based on a multivariate Breiman Lemma is exposed for building multivariate regularly varying distributions. It encompasses the classical max-stable models such as logistic, negative logistic, Dirichlet, Husler-Reiss, extremal-Gaussian and extremal-t. In a second part, we focus on the Pareto model associated with the Husler-Reiss max-stable model. An interesting exponential family property is exhibited and maximum-likelihood estimation is considered. The case of different tail indices for different margins is also studied.

E0710: **Efficient simulation of Brown-Resnick processes by means of locally equivalent log-Gaussian representations**
Presenter: **Kirstin Strokorb**, Cardiff University, United Kingdom
Co-authors: Marco Oesting
Among max-stable processes, the class of Brown-Resnick processes seems particularly attractive for practical use. Several algorithms have been suggested for its simulation. While the efficiency of such algorithms can be viewed as a function of the number of points on which the process is simulated, it is more difficult to understand the role of the simulation domain in which they lie, on which we focus in this project. Our suggestion improves the efficiency and/or accuracy of a stopping-time based algorithm. We show that a relatively simple and non-costly adjustment of previous algorithms can be very beneficial.

EO657 Room MAL B30 FUNCTIONAL DATA ANALYSIS, METHODS AND APPLICATIONS Chair: Donatello Telesca

E0307: **Hybrid principal components analysis for region-referenced longitudinal functional EEG data**
Presenter: **Damla Senturk**, University of California Los Angeles, United States
Co-authors: Aaron Scheffler, Donatello Telesca, Qian Li, Catherine Sugar, Charlotte DiStefano, Shafali Jeste
The electroencephalography (EEG) data produce data frames of complex structure that includes functional, longitudinal, and regional dimensions. Our motivating example is a word segmentation paradigm where typically developing (TD) children and children with Autism Spectrum Disorder (ASD) were exposed to a continuous speech stream. For each subject, continuous EEG signals recorded were transformed into the frequency domain resulting in region-referenced principal power where one-second segments throughout the experiment represent the longitudinal dimension, principal power obtained across frequencies represent the functional dimension and the scalp regions represent the regional dimension. We propose a hybrid principal components analysis (HPCA) for region-referenced longitudinal functional EEG data which utilizes both vector and functional principal components analyses and does not collapse information along any of the three dimensions of the data. The proposed decomposition only assumes weak separability of the higher-dimensional covariance process and utilizes a product of one dimensional eigenvectors and eigenfunctions,

obtained from the frequency, segment, and region marginal covariances, to represent the observed data, providing a computationally feasible nonparametric decomposition. A mixed effects modeling framework is proposed to estimate the model components, coupled with a bootstrap test for group level inference; both geared towards sparse data applications.

E0648: **Multiplicative component models for replicated point processes**
Presenter: **Daniel Gervini**, University of Wisconsin-Milwaukee, United States
A semiparametric approach is presented to the analysis of replicated temporal and spatial point processes, based on a multiplicative decomposition of the intensity functions. The method applies to univariate and multivariate processes, and in the latter case we discuss dimension reduction techniques. We illustrate this approach with the joint spatial analysis of various types of crimes in the city of Chicago.

E0655: **A robust t-process regression model with independent errors**
Presenter: **Jian Qing Shi**, Newcastle University, United Kingdom
Gaussian process regression (GPR) models are well known to be susceptible to outliers. Robust process regression models based on t process or other heavy tailed processes have been developed to address the problem. However, due to the nature of the current definition for heavy-tailed processes, the unknown process regression function and the random errors are always defined jointly and thus dependently. This definition, mainly owing to the dependence assumption involved, is not justified in many practical problems and thus limits the application of those robust approaches. It also results in a limitation of the theory of robust analysis. We will discuss a new robust process regression model enabling independent random errors and will also discuss an efficient estimation procedure. We will present an application to analyse medical game data and show that the proposed method is robust against outliers and has a better performance in prediction compared with the existing models.

E0676: **Joint sparse curves clustering and alignment**
Presenter: **Valeria Vitelli**, University of Oslo, Norway
The problem of curve clustering in presence of misalingment is considered. This is a frequent situation when dealing with functional data. A method to jointly cluster and align curves, which efficiently decouples amplitude and phase variability by detecting amplitude clusters while simultaneously disclosing clustering structures in the phase, is described, and its efficacy is demonstrated on a couple of real applications. On the other hand, finding sparse solutions to clustering problems has emerged as a hot topic in statistics in recent years, due to the technological improvements in measurement systems leading to the spread of high-dimensional data in many real applications. This problem has not yet been properly treated in the literature on functional data, even though it is often of much interest to select the curves' most relevant features while jointly solving a classification problem. Functional sparse clustering can be analytically defined as a variational problem with a hard thresholding constraint ensuring the sparsity of the solution: this problem is shown to be well-posed, to have a unique optimal solution, and to provide good insights in real applications. Finally, a possible approach to deal with sparse functional clustering when curves are misaligned is also proposed, and some preliminary results are shown.

EO144 Room MAL B35 STATISTICS IN NEUROSCIENCE Chair: Jeff Goldsmith

E0313: **Inter-modal coupling changes through development and in neuropathology**
Presenter: **Russell Shinohara**, University of Pennsylvania Perelman School of Medicine, United States
A proliferation of MRI-based neuroimaging modalities now allows measurement of diverse features of brain structure, function, and connectivity during the critical period of adolescent brain development. However, the vast majority of developmental imaging studies use data from each neuroimaging modality independently. As such, most developmental studies have not considered, or have been unable to consider, potentially rich information regarding relationships between imaging phenotypes. At present, it remains unknown how local patterns of structure and function are related, how this relationship changes through adolescence as part of brain development, and how developmental pathology may impact such relationships. We propose to measure the relationships between measures of brain structure, function, and connectivity during adolescent brain development by developing novel, robust analytic tools for describing relationships among imaging phenotypes. Our over-arching hypothesis is that such relationships between imaging features will provide uniquely informative data regarding brain health, over and above the content of data from each modality when considered in isolation.

E0849: **Modeling motor learning using heteroskedastic functional principal components analysis**
Presenter: **Jeff Goldsmith**, Columbia University, United States
Co-authors: Daniel Backenroth, Michelle Harran, Juan Cortes, John Krakauer, Tomoko Kitago
Experiments involving kinematic data – dense recordings of hand or finger position over time during the execution of a motion – provide deep insights into the processes underlying motor control and learning. To model the reduction of motion variance achieved through repetition, we extend the functional principal components analysis framework to allow subject and covariate effects on score variances. In a setting where the components are invariant across subjects and covariate values, this approach provides a flexible and interpretable way to explore factors that affect the variability of functional data. Parameters are jointly estimated in a Bayesian framework using both MCMC and a computationally efficient variational approximation.

E1406: **Spectral causality: Exploring lead-lag dependence structure between oscillatory activities in multivariate signals**
Presenter: **Hernando Ombao**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Co-authors: Abdulrahman Althobaiti
The motivation comes from the problem of characterizing multi-scale changes in brain signals following an event (e.g., stroke, epileptic seizure). Preliminary analyses of brain signals following stroke show that there are both short-term (immediate) responses to stroke and long term as affected neuronal populations undergo a reorganization in response to an injury. Spectral analysis will be used to study dependence between neuronal populations. Of prime interest is the notion of "spectral causality", which is broadly characterized as the extent to which an oscillatory activity in a population of neurons can predict various oscillatory activities in another region at a future time point. Our approach is to extract different oscillatory components via linear filtering and then examine cross-dependence between the filtered signals. The proposed spectral causality approach overcomes the limitations of classical measures such as coherence and partial coherence since these do not indicate directionality. In addition, the proposed approach is superior to partial directed coherence because it is able to precisely capture the time lag in the between oscillatory activity at different regions. Interesting results from exploratory analyses, showing the immediate changes and long-term brain response, will be reported.

E1496: **Challenges and opportunities in population neuroimaging**
Presenter: **Thomas Nichols**, University of Oxford, United Kingdom
Brain imaging studies have traditionally struggled to break into 3-digit sample sizes: e.g., a recent Functional Magnetic Resonance Imaging (fMRI) meta-analysis of emotion found a median sample size of $n = 13$. However, we now have a growing collection studies with sample sizes with 4-, 5- and even 6-digits. Many of these 'population neuroimaging' studies are epidemiological in nature, trying to characterise typical variation in the population to help predict health outcomes across the life span. We will discuss some of the challenges these studies present, in terms of massive computational burden but also in ways that they expose shortcomings of existing mass univariate techniques. We will also discuss how these

datasets present intriguing methodological problems heretofore absent from neuroimaging statistics. For example, the 'null hypothesis fallacy' is how H_0 is never strictly true, and yet with 100,000 subjects you will eventually find some effect even if it is meaningless. This motivates work spatial confidence sets on meaningful effect sizes (instead of thresholding test statistic images), providing intuitive measures of spatial uncertainty. We will discuss these findings and other work our group had done on dealing with these problems and sizing the opportunities.

EO248 Room Woburn STATISTICAL METHODS FOR BUSINESS AND FINANCE Chair: Tim Verdonck

E1340: **Monetary policy transmission under changing heterogeneous population**
Presenter: **Ken Bastiaensen**, UGent, Belgium

Secular stagnation currently constitutes one of the main debates in economics: Low economic growth, low inflation and low interest rates have become commonplace in many developed economies. One potential driving factor of secular stagnation is that monetary policy is, in the presence of ageing, no longer an effective tool to counter a deficient demand for investment goods in the economy. The focus is on saving and investment (S&I) decisions given the heterogeneity of individuals with regard to age, expected life expectancy, civil state, income level, income stability, wealth, indebtedness, spending patterns, etcetera. We use several tools from Statistical Learning and more traditional econometrics models such as local projection modelling.With this approach we are able to gain insight into past behaviour, uncovering patterns in groups of individuals that are not obvious from aggregate observables. More importantly, we are able to predict future S&I decisions for each individual, under given future scenarios of macroeconomic conditions such as interest rate. Special care has been given to avoid overfitting. The behaviour arises solely from the experience of other agents without solving any structural economic model as assumed in traditional economics. Our approach has been employed with remarkable out-of-sample success (1 year ex post) on millions of real-world agents. Preliminary results show changing (non-linear) behaviour with regard to changing characteristics.

E0389: **Profit driven decision trees for churn prediction**
Presenter: **Sebastiaan Hoeppner**, KU Leuven, Belgium
Co-authors: Eugen Stripling, Bart Baesens, Seppe vanden Broucke, Tim Verdonck

The interest for data mining techniques has increased tremendously during the past decades, and numerous classification techniques have been applied in a wide range of business applications. In the telecommunication sector, companies heavily rely on predictive churn models to detect churners in a vast customer base and to remain competitive in a saturated market. In a recent paper, the expected maximum profit (EMP) has been proposed which explicitly takes the cost of offer and the customer lifetime value of retained customers into account. It thus permits the selection of the most profitable classifier which better aligns with business requirements of end-users and stake holders. However, modelers are currently limited to applying this metric in the evaluation step. Therefore, we present a classifier named ProfTree, that maximizes the EMP metric in the training step using a genetic algorithm. The technique is based on a classification tree for modeling a binary response variable.

E0453: **Cellwise robust M estimation based on sparse outlyingness**
Presenter: **Sven Serneels**, BASF Corporation, United States
Co-authors: Tim Verdonck, Sebastiaan Hoppner

Robust statistical estimators have two major practical purposes: stable estimation in the presence of outliers and outlier detection. Outliers are considered to be entire cases in a sample, regardless of dimensionality. State-of-the-art outlier detection methods based on robust statistics therefore flag entire cases as outliers. However, as data dimensions increase, it becomes increasingly more likely that outliers only deviate with respect to a subset of the variables that constitute them. It has recently been shown that the direction of maximal oultyingness can be rewritten as a regression problem. By applying a variable selection technique to that associated regression problem, those variables that contribute most to a case's outlyingness, can be detected. This detection scheme can be iterated until convergence, stopping when the case is no longer outlying after removal of those variables that contributed most to its outlyingness. This information can be embedded into a robust estimation procedure. It is well known that M estimators can efficiently be implemented in an iterative re-weighting scheme. Given the information on individual variables' contribution to outlyingness, the iterative re-weighting scheme can be adapted to use cell specific weights instead of case weights. The estimator thus constructed, is a cellwise robust M estimator. A few examples of how this general idea can be implemented to specific estimators, will be shown.

E0820: **Shrinkage approaches for the estimation of higher order comoments**
Presenter: **Dries Cornilly**, Vrije Universiteit Brussel, Belgium
Co-authors: Kris Boudt, Tim Verdonck

Optimal financial decision making often requires to take into account the higher order comoments of dependent random variables. They are however notoriously difficult to estimate. We propose several shrinkage approaches that combine the sample estimator with a structured target comoment estimate that are optimal according to a minimum distance criterion. The first contribution is in terms of the coskewness matrix where we propose unbiased estimates for the MSE loss function determining the shrinkage intensity of the multi-target shrinkage estimator. The second contribution is to obtain estimates for the higher order comoments under a latent factor model. Under this approach, we estimate the higher order comoments structurally by means of minimum distance estimation, in order to incorporate the underlying factor structure. Asymptotic normality of the estimator is proven. In both cases, the good properties are confirmed in simulations and applications on hedge fund return data.

EO528 Room MAL 402 SUFFICIENT DIMENSION REDUCTION AND INVERSE REGRESSION Chair: Xin Zhang

E0265: **The robustness of high dimensional methods to model misspecification: A dimension reduction tale**
Presenter: **Matey Neykov**, Carnegie Mellon University, United States

Recently there has been a flurry of activity in the compressed sensing and statistics communities to study robust estimation and variable selection of popular high dimensional regularized procedures. It has been argued that LASSO achieves optimal statistical rates over a broad class of models including linear regression as a special case. This line of thinking will be motivated via classical results established in the dimension reduction literature. Examples to corroborate these ideas will be provided. In particular, we will discuss the robustness of LASSOs support recovery over single index models with Gaussian design, and will argue that the thresholded Wirtinger Flow algorithm is applicable beyond the real phase retrieval model.

E0503: **Model-free variable selection for the regression mean**
Presenter: **Yuexiao Dong**, Temple University, United States

A novel test statistic is proposed to identify important predictors for the conditional mean function in regression. The stepwise regression algorithm based on the proposed test statistic guarantees variable selection consistency without specifying the functional form of the conditional mean. When the predictors are ultrahigh dimensional, a model-free screening procedure is introduced to precede the stepwise regression algorithm. The screening procedure has the sure screening property when the number of predictors grows at an exponential rate of the available sample size. The finite-sample performances of our proposals are demonstrated via numerical studies.

E1334: **Copula Gaussian graphical models for functional data**
Presenter: **Bing Li**, The Pennsylvania State University, United States

The problem of constructing statistical graphical models for functional data is considered; that is, the observations on the vertices are random functions. This types of data are common in medical applications such as EEG and fMRI. Recently published functional graphical models rely on the assumption that the random functions are Hilbert-space-valued Gaussian random elements. We relax this assumption by introducing a copula Gaussian random elements Hilbert spaces, leading to what we call the Functional Copula Gaussian Graphical Model (FCGGM). This model removes the marginal Gaussian assumption but retains the simplicity of the Gaussian dependence structure, which is particularly attractive for large data. We develop four estimators, together with their implementation algorithms, for the FCGGM. We establish the consistency and the convergence rates of one of the estimators under different sets of sufficient conditions with varying strengths. We compare our FCGGM with the existing functional Gaussian graphical model by simulation, under both non-Gaussian and Gaussian graphical models, and apply our method to an EEG data set to construct brain networks.

E1407: **Biomarker discovery in heterogeneous populations**
Presenter: **Elizabeth Slate**, Florida State University, United States
Co-authors: Junxian Geng

Identification of valid, clinically relevant biomarkers for disease has potential to provide less invasive diagnostic tools, to enhance understanding of initiation and progression at the cellular level, and to guide development of new therapeutic agents. When the biomarkers are binary, logic regression provides a means to discover Boolean combinations of the markers strongly associated with outcome. The interpretability of these Boolean marker combinations and, potentially, additional interactions with environmental and behavioral characteristics, is appealing and can provide insight. However, complex diseases such as cancer that arise from multiple pathways and present at varying stages of development and progression can lead to hidden population heterogeneity in the biomarker-disease association. We describe an extension of logic regression for jointly modeling binary and continuous outcomes that uses a latent class structure to accommodate subpopulation heterogeneity. Estimation and inference are compared for two Bayesian semiparametric formulations using a variety of computational approaches.

EO378 Room MAL 414 ADVANCED METHODS IN BIOSTATISTICS Chair: Yuko Araki

E1216: **Functional path analysis with composite basis expansions**
Presenter: **Yuko Araki**, Shizuoka University, Japan

Recent years have seen that functional data analysis are capable of extracting intrinsic features from recently arising complicated and high dimensional data, such as three dimensional brain sMRI, time course microarray data, or hundreds of records of human gait, for example. We introduce statistical methods for describing the direct and indirect dependencies among a set of variables including high dimensional covariates. The proposed method is based on composite basis function, which is an extended version of basis expansions with the help of sparse PCA. The proposed models are applied to real data example and Monte Carlo simulations are conducted to examine the efficiency of our modelling strategies.

E0936: **Identifiability and multicollinearity in scalar-on-functions regression**
Presenter: **Clara Happ**, LMU Munich, Germany
Co-authors: Sonja Greven

Data with functional features arises in more and more disciplines, bringing new possibilities and challenges to practitioners and statisticians. One key method in functional data analysis is modeling the relationship of one or more functional predictor variables and a scalar response, which is referred to as scalar-on-functions regression. As for all regression models, there is need for identifiable coefficient functions as a necessary condition for obtaining interpretable results. The issue of non-identifiability of scalar-on-function terms and its natural extension, functional multicollinearity, is discussed for the functional linear model. Starting from a theoretical point of view, practical diagnostic criteria are developed and countermeasures are proposed for the two main estimation approaches, penalized scalar-on-function regression and functional principal component regression. The theoretical results are verified in an extensive simulation study and their relevance is illustrated in an application to biomedical data.

E0661: **Time-dependent summary receiver operating characteristics for meta-analysis of prognostic studies**
Presenter: **Satoshi Hattori**, Osaka University, Japan

Prognostic studies are widely conducted to examine whether biomarkers are associated with patient's prognoses and play important roles in medical decisions. Since findings from one prognostic study may be very limited, meta-analyses may be useful to obtain sound evidence. However, prognostic studies are often analyzed by relying on a study-specific cut-off value, which can lead to difficulty in applying the standard meta-analysis techniques. We propose two methods to estimate a time-dependent version of the summary receiver operating characteristics curve for meta-analyses of prognostic studies with a right-censored time-to-event outcome. We introduce a bivariate normal model for the pair of time-dependent sensitivity and specificity, and propose a method to form inferences based on summary statistics reported in published papers. This method provides a valid inference asymptotically. In addition, we consider a bivariate binomial model. To draw inferences from this bivariate binomial model, we introduce a multiple imputation method. The multiple imputation is found to be approximately proper multiple imputation and thus the standard Rubin's variance formula is justified from a Bayesian view point. Our simulation study and application to a real dataset revealed that both methods work well with a moderate or large number of studies and the bivariate binomial model coupled with the multiple imputation outperforms the bivariate normal model with a small number of studies.

E0914: **Estimation of well-clustered structure via penalized maximum likelihood method in factor analysis model**
Presenter: **Kei Hirose**, Kyushu University, Japan
Co-authors: Yoshikazu Terada

A prenet (PRoduct Elastic NET) penalization is proposed to estimate a well-clustered structure in a factor analysis model. The penalty is constructed by the product of a pair of parameters in each row of the loading matrix. A remarkable feature of the prenet is that a large amount of penalization leads to the perfect simple structure, which is completely well-clustered. Furthermore, the perfect simple structure estimation via the prenet penalty is shown to be a generalization of the k-means variables clustering. On the other hand, with a mild amount of prenet penalization, the estimated loading matrix is approximated by that obtained using the quartimin rotation, a widely used oblique rotation method. The proposed procedure is available for use in the R package fanc, which is available at http://cran.r-project.org/web/packages/fanc.

EO629 Room MAL 415 RECENT DEVELOPMENTS IN HYPOTHESIS TESTING FOR HIGH DIMENSIONAL DATA Chair: Ping-Shou Zhong

E1423: **Tests for coefficients in high dimensional heteroscedastic linear models**
Presenter: **Ping-Shou Zhong**, Michigan State University, United States
Co-authors: Honglang Wang, Yuehua Cui
Hypothesis testing problems are considered for a low-dimensional coefficient vector in a high-dimensional linear model with heteroscedastic variance. Heteroscedasticity is a commonly observed phenomenon in many applications including finance and genomic studies. Several statistical inference procedures have been proposed for low-dimensional coefficients in a high-dimensional linear model with homoscedastic variance. However, existing procedures designed for homoscedastic variance are not applicable for models with heteroscedastic variance and the heterscedasticity issue has been rarely investigated and studied. We propose a simple inference procedure based on empirical likelihood to overcome the heteroscedasticity issue. The proposed method is able to make valid inference even when the conditional variance of random error is an unknown function of high-dimensional predictors. We apply our inference procedure to three recently proposed estimating equations and establish the asymptotic distributions of the proposed methods. Simulation studies and real data analyses are conducted to demonstrate the proposed methods.

E1427: **Detecting adverse drug reactions from pharmacovigilance databases**
Presenter: **Kun Liang**, University of Waterloo, Canada
The World Health Organization and many countries have built pharmacovigilance databases to detect potential adverse reactions due to marketed drugs. Although a number of methods have been developed for early detection of adverse drug effects, the vast majority of them do not consider the multiplicity arising from testing thousands drug and adverse event combinations. We first derive the optimal statistic to maximize the power of detection while maintaining proper error rate. We then propose a nonparametric empirical Bayes method to estimate the optimal statistic and demonstrate its superior performance through simulation. Finally, the proposed method is applied to the pharmacovigilance database in the United Kingdom.

E1448: **Dependence tests in high-dimensional settings under a Kronecker product covariance decomposition**
Presenter: **Anestis Touloumis**, University of Brighton, United Kingdom
In many applications with high-dimensional data, the subject-specific data can be considered as realizations of large matrix-variate variables where both the rows and the columns correspond to features of interest and dependencies might occur among and between the row and column variables. For example, consider multi-tissue studies in genetics where for each subject gene expression levels are measured in multiple tissues. The subject-specific data can be written in a matrix form where the row variables correspond to genes, the column variables to tissues and the expression levels are the measurements. For inferential purposes, researchers usually employ a Kronecker product form of two covariance matrices for the dependence structure between the row and column variables; one that describes the dependence structure among the row variables and the other describes the dependence structure among the column variables. However, there is a lack of hypothesis testing procedures for the covariance matrices in high-dimensional settings. We present non-parametric tests for the identity, sphericity and diagonal sphericity hypothesis for the row (column) covariance matrix while treating the column (row) covariance matrix as nuisance.

E1456: **Simultaneous confidence bounds for the false discovery proportion: A permutation approach**
Presenter: **Aldo Solari**, University of Milano-Bicocca, Italy
Co-authors: Jesse Hemerik, Jelle Goeman
Testing of multiple hypotheses has received much attention in recent years due to the availability of ever-larger data sets. It is often desirable to reject as many hypotheses as possible while keeping the false discovery proportion (FDP) in check. The FDP attracted considerable interest recently because under strong dependence among p-values, it represents a more relevant measure than the False Discovery Rate (FDR), which is the expected value of the FDP. In fact, the stronger the positive dependence among p-values, the higher the variability of FDP around its expected value, and the FDR can be far from the actual FDP. However, high dimensionality and dependency impose a formidable methodological challenge in constructing tight upper bounds for the FDP. By using a permutation approach. A powerful method to derive simultaneous confidence bounds for the FDP has been previously provided. We show that such a method can be generalised and improved. First, we allow for more flexibility in constructing the simultaneous confidence bounds. Second, we show how the power of the previous method can be uniformly increased by using closed testing.

EO168 Room MAL 416 DISCRETE-VALUED PROCESSES AND DYNAMIC NETWORKS Chair: Carsten Jentsch

E0477: **Time series modeling on dynamic networks**
Presenter: **Jonas Krampe**, TU Braunschweig, Germany
Multivariate time series on dynamic networks with a fixed number of vertices are considered. Each component of the time series is assigned to a vertex of the underlying network. The dynamic dependency between the various components of the time series is modeled by means of the edges. We make use of a multivariate doubly stochastic time series framework, that is we assume linear processes for which the coefficient matrices are stochastic processes themselves. We explicitly allow for dependence in the dynamics of the coefficient matrices, including of course an i.i.d. structure as is typically assumed in random coefficients models. Conditions for stationarity will be given and asymptotic normality of simple statistics like the sample mean is investigated. Furthermore, autoregressive moving average models are defined in this framework. Estimators of the parameters are discussed and how this can be used to forecast such a process. Some interesting features of these processes are shown in simulations and the finite sample behavior of the forecast approach is investigated.

E0463: **Nonparametric inference for continuous-time event counting and link-based dynamic network models**
Presenter: **Alexander Kreiss**, Heidelberg University, Germany
Co-authors: Enno Mammen, Wolfgang Polonik
A flexible approach for modeling both dynamic event counting and dynamic link-based networks based on counting processes is proposed, and estimation in these models is studied. We consider nonparametric likelihood based estimation of parameter functions via kernel smoothing. The asymptotic behavior of these estimators is rigorously analyzed by allowing the number of nodes to tend to infinity. The finite sample performance of the estimators is illustrated through an empirical analysis of bike share data.

E1145: **Generalized binary time series models**
Presenter: **Lena Reichmann**, University of Mannheim, Germany
Co-authors: Carsten Jentsch
The serial dependence of categorical data is commonly described using Markovian models. As a result of their flexibility, they can suffer from a huge number of parameters if the state space or the model order becomes large. To address the problem of a large number of model parameters in the univariate case, a more parsimonious and nicely interpretable class of Discrete AutoRegressive Moving-Average (DARMA) models has been introduced in the literature. For binary data, we propose two model extensions. First, we allow for negative model coefficients to allow also for negative autocorrelations, which is not possible using DARMA models. Second, we consider the vector-valued case that is suitable e.g. for

dynamic network modeling. In this case, the effect of Markov models having a huge number of parameters becomes even more pronounced. Both extensions are simple and maintain the nice interpretability and the autoregressive moving-average structure leading to a new generalized DARMA model class. We provide sufficient stationarity conditions and derive the stationary solution of the model equations. For the purely autoregressive case, we prove Yule-Walker-type equations that facilitate the task of parameter estimation in these models to some large extent. Further, we discuss mixing properties of these models. For illustration, we study the estimation performance in these models by simulations and apply our model to quarterly OECD recession data from G7 countries.

E1001: **Modeling and prediction of dynamic networks using binary autoregressive time series processes**
Presenter: **Carsten Jentsch**, University of Mannheim, Germany
Co-authors: Lena Reichmann, Shaikh Tanvir Hossain

Suppose a time series of networks is identified by their adjacency matrices $A_1,\ldots,A_T$, where $A_t = (a_{ij;t})_{i,j=1,\ldots,N}$ with $a_{ij;t} \in \{0,1\}$ and $a_{ij,t} = 1$ indicating that there is a directed edge pointing from vertex i to vertex j at time t. To model the joint dynamics of the edges, we propose to use multivariate binary time series processes. For this purpose, we adopt the class of Discrete AutoRegressive Moving-Average (DARMA) models for univariate categorical data. Recent extensions of these models allow the application to vector-valued data and to model negative autocorrelations by a simple modification. The resulting model class is flexible enough to capture very general autocorrelations driving the dynamic network structure. For the purely autoregressive case, Yule-Walker-type equations hold that allow in principle an explicit estimation of all model parameters. However, as the dimension of the adjacency matrices grows quadratically with the number of vertices, we shall make use of Lasso-penalization techniques to estimate sparse models. For this purpose, we adopt a previous approach which provides with consistent estimators for high-dimensional vector autoregressive models under sparsity. Our modeling approach is suitable for prediction of single and joint edge probabilities in dynamic networks. We illustrate our method by simulations and for real data.

EO330 Room MAL 421 ADVANCES IN ROBUST STATISTICAL LEARNING Chair: Gen Li

E0325: **A general framework for association analysis of heterogeneous data**
Presenter: **Gen Li**, Columbia University, United States
Co-authors: Irina Gaynanova

Multivariate association analysis is of primary interest in many applications. Despite the prevalence of high-dimensional and non-Gaussian data (such as count-valued or binary), most existing methods only apply to low-dimensional datasets with continuous measurements. Motivated by the Computer Audition Lab 500-song (CAL500) music annotation study, we develop a new framework for the association analysis of two sets of high-dimensional and heterogeneous (continuous/binary/count) data. We model heterogeneous random variables using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to identify shared and individual patterns for two datasets. We also introduce a new measure of the strength of association, and a permutation-based procedure to test its significance. An alternating iteratively reweighted least squares algorithm is devised for model fitting, and several variants are developed to expedite computation and achieve variable selection. The application to the CAL500 data sheds light on the relationship between acoustic features and semantic annotations, and provides an effective means for automatic annotation and music retrieval.

E0372: **Discrete smoothing kernels**
Presenter: **Marianthi Markatou**, University at Buffalo, United States

Kernels are essential elements in the construction of learning systems and have received considerable attention in the machine learning literature. In statistics, kernels are used as tools for achieving specific data analytic goals such as density estimation or goodness of fit testing. We consider the problem of creating smoothing kernels for multinomial and product multinomial models. Our construction is based on the properties of continuous time Markov chains. We will discuss an algorithm for the construction of these kernels and exemplify its use in smoothing ordered categorical data. Furthermore, using these constructions, we will discuss independence tests that are analogues of the chi-squared test of independence.

E0326: **Multicategory classification via forward-backward support vector machine**
Presenter: **Donglin Zeng**, University of North Carolina at Chapel Hill, United States
Co-authors: Yuanjia Wang, Xuan Zhou

A novel and computationally efficient learning algorithm, namely forward-backward support vector machine (FB-SVM), is proposed to perform multicategory learning. The new method is based on a sequential binary classification algorithms: we first classify a particular class by excluding the possibility of labeling as any other classes using a forward procedure of sequential SVM; we then exclude already classified classes and repeat the same learning for the remaining classes in a backward way. The proposed algorithm relies on support vector machines for each binary classification and utilizes only feasible data in each step of this procedure; therefore, the method guarantees convergence and computation burden is little. Furthermore, we show that the derived rule from FB-SVM is Fisher consistent and we obtain the risk bound for the predicted misclassification rate. We conduct extensive simulation studies, including benchmark examples from existing methods, to demonstrate that the proposed method has superior performance in terms of small misclassification rates and significantly improved computation speed. Finally, we apply the proposed method to analyze real data for further illustration.

E0741: **Spatially weighted reduced-rank framework for functional MRI data**
Presenter: **Mihye Ahn**, University of Nevada Reno, United States

Recently, the analysis of functional imaging data has received much attention to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.

EO433 Room MAL 532 STATISTICAL CHALLENGES FOR SPATIAL EPIDEMIOLOGICAL DATA Chair: Veronica Berrocal

E1244: **Uncertainty in pollutant concentration surface under preferential sampling**
Presenter: **Annibale Biggeri**, University of Florence, Italy
The focus is on environmental statistics, with the aim of estimating the concentration surface and related uncertainty of an air pollutant. We used air quality data recorded by a network of monitoring stations within a Bayesian framework to overcome difficulties in accounting for prediction uncertainty and to integrate information provided by deterministic models based on emissions meteorology and chemico-physical characteristics of the atmosphere. We considered the situation in which the spatial process of interest and the sampling locations are not independent. This is known in the literature as the preferential sampling problem. We developed a Bayesian geostatistical model to account for preferential sampling, statistical integration and uncertainty. We used PM10 data arising from the air quality network of the Environmental Protection Agency of Lombardy Region (Italy) and numerical outputs from the deterministic model. We specified an inhomogeneous Poisson process for the sampling locations intensities and a shared spatial random component model for the dependence between the spatial location of monitors and the pollution surface. We found greater predicted standard deviation differences in areas not properly covered by the air quality network. In conclusion, in this context inferences on prediction uncertainty may be misleading when geostatistical modelling does not take into account preferential sampling.

E0715: **Latent space models for ecological networks**
Presenter: **Catherine Calder**, The Ohio State University, United States
Co-authors: Christopher Browning
Research on neighborhood effects often focuses on linking features of social contexts or exposures to health, educational, and criminological outcomes. Traditionally, individuals are assigned a specific neighborhood, frequently operationalized by the census tract of residence, which may not contain the locations of routine activities. In order to better characterize the many social contexts to which individuals are exposed as a result of the spatially-distributed locations of their routine activities and to understand the consequences of these socio-spatial exposures, we have developed the concept of ecological networks. Ecological networks are two-mode networks that indirectly link individuals through the spatial overlap in their routine activities. The focus is on statistical methodology for understanding the structure underlying ecological network. In particular, we propose a continuous latent space model that allows for third-order dependence patterns in the interactions between individuals and the places they visit. We illustrate our methodology using activity pattern and sample survey data from Columbus, OH.

E1023: **Spatio-temporal models to evaluate the effect of extreme hydro-climatic events on birth wight**
Presenter: **Erick Chacon Montalvan**, Lancaster University, United Kingdom
Newborn weight is used as an indicator of population health because it is related to long-term outcomes in education, income and disability. There are a number of existing studies that have been conducted to identify the major determinants of birth weight. We consider how the environment can impact on newborn health outcomes. In particular, the goal is to evaluate the effects of exposure to floods and droughts during pregnancy on birth weight in roadless cities of the Brazilian Amazonas state. We make use of generalized additive models for location, scale and shape and divide the process into two stages. First, we propose a new model-based standardized precipitation index to quantify the magnitude of floods and droughts. Then, this index is used as covariate to evaluate the effects of exposure to extreme hydro-climatic events. The results suggest: birth weight has a seasonal behavior with respect to river levels; there is a global negative trend on birth weight; there exist vulnerable groups; and that exposure to extreme hydro-climatic events can have a negative impact on birth weight.

E0874: **Spatiotemporal modeling of vector-borne disease risk**
Presenter: **Gavino Puggioni**, University of Rhode Island, United States
The occurrence of mosquito-borne tropical diseases, such as Dengue and Zika, have been rising in the last ten years and linked to changes in precipitation, temperature, and urbanization. An illustrative case study features monthly Dengue reports at the municipal level in Puerto Rico from 1990 to 2015, weather variables collected from 34 stations around the island, and satellite data. The first stage of the proposed modeling strategy addresses the difference in spatial support of predictors and response. At the second stage, several space-time CAR specifications are implemented in a Bayesian framework to assess the relative risk of these factors. The model's predictive framework can be used to inform early warning systems for targeted surveillance and outbreak detection.

EO437 Room MAL 538 BAYESIAN MODELS FOR CLUSTERING Chair: Maria De Iorio

E1047: **Bayesian nonparametric covariate-driven clustering: An application to blood donors data**
Presenter: **Raffaele Argiento**, University of Torino, Italy
Blood is an important resource in global healthcare and therefore an efficient blood supply chain is required. Predicting arrivals of blood donors is fundamental since it allows for better planning of donations sessions. With the goal of characterizing behaviors of donors, we analyze gap times between consecutive blood donations. In order to take into account population heterogeneity we adopt a Bayesian model for clustering. In such a context, defining the model boils down to assign the prior for the random partition itself and to flexibly assign the cluster-specific distribution, since, conditionally on the partition, data are assumed iid within each cluster and independent between clusters. In particular, we drive the prior knowledge on the random partition by increasing the probability that two donors with similar covariates belong to the same cluster. The resulting model is a covariate-dependent nonparametric prior, thus departing from the standard exchangeable assumption. Specifically, we modify the prior on the partition prescribed by the class of normalized completely random measures by including in the prior a term that takes into account the distance between covariates. First, briefly discuss the theoretical implications of this mathematical operation, finally, we fit our model to a large dataset provided by the Milan department of AVIS (Italian Volunteer Blood-donors Association) the largest provider of blood donations in Italy.

E1018: **A hierarchical nonparametric approach for robust graphical modelling**
Presenter: **Andrea Cremaschi**, Universitetet i Oslo, Norway
Co-authors: Raffaele Argiento
Useful tools for exploring multivariate network structures are Gaussian graphical models. However, alternative models are needed when data are strongly non-Gaussian. The t-Student distribution, obtained by dividing each component of the data vector by a gamma random variable, is the straightforward generalisation to accommodate such issue. The Dirichlet t-Student distribution is obtained when the law of the divisors is the Dirichlet process. In the latter, conditionally to a shared mass parameter, a Dirichlet process is introduced for every multivariate observation, so that one can cluster the components of each data point according to their deviation from the Normal distribution (outlier clustering). We consider a more general class of nonparametric distributions, namely the class of normalised completely random measures (NormCRM), which yields a more flexible component clustering. Moreover, in order to borrow more information across the data, we model the dependence among the NormCRM through a nonparametric hierarchical structure. At data level, each NormCRM is centred on the same base measure, which is a NormCRM itself. The discreteness of the shared base measure implies that the processes at data level share the same atoms. This desired feature allows to cluster together components of different data.An application to a bio-medical dataset is described for illustrative purposes.

E1080: **Bayesian robustness in product partition models**
Presenter: **Rosangela Loschi**, Universidade Federal de Minas Gerais, Brazil
Co-authors: Jacqueline Alves Ferreira, Fabrizio Ruggeri
The focus is on the robustness analysis of non-exchangeable product partition models (PPM), which are widely used for multiple change points detection. Bayesian robustness is usually concerned with the impact of perturbations in the prior distribution of the parameter of interest on its posterior inference. We consider multiplicative perturbations in the data distribution, as well as in the prior distribution of its parameters. As a novelty in the robust Bayesian and PPM literature, we introduce some sensitivity measures to examine how those perturbations are affecting the posterior inference about the number of clusters and their position, as well as the product estimates. We focus our analysis on the skew-normal class of distributions, thus building a PPM under skew-normality. We apply the proposed PPM to analyze a Brazilian tomato price data set.

E1157: **Accommodating missing covariates via product partition models**
Presenter: **Garritt Page**, Brigham Young University, United States
Co-authors: Fernando Quintana
Missing observations are ubiquitous in data driven research studies. Statistical methods that accommodate different types of missingness have been developed and now have a long history. Many of these methods depend on assumptions about the missingness mechanism (e.g., missing completely at random) that are difficult to verify in a data driven way. The default method of choice (particularly in a Bayesian setting) is to assume missing at random and carry out multiple imputation where values for missing observations are imputed and analysis are run on the filled in data set. To avoid making it difficult to verify assumptions about the missing mechinism, we develop a methodology that flexibly accommodates missingness based on covariate dependent random partition models. Without discarding subjects with partially observed covariate vectors, we simply employ those variables that have been measured when forming clusters and making predictions. We apply the methodology to data gathered in a study of osteonecrosis.

EO306 Room MAL 539 LATENT VARIABLE MODELS WITH APPLICATIONS Chair: Sara Taskinen

E0536: **Comparing estimation methods for generalized linear latent variable models**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Francis Hui, Jenni Niku, David Warton
In many studies in community ecology, multivariate abundance data are often collected. Such data are characterized by two main features. First, the data are high-dimensional in that the number of species often exceeds the number of sites. Second, the data almost always cannot be suitably transformed to be normally distributed. Instead, the most common types of responses recorded include presence-absence records, overdispersed species counts, biomass, and heavily discretized percent cover data. One promising approach for modelling data described above is generalized linear latent variable models. By extending the standard generalized linear modelling framework to include latent variables, we can account for covariation between species not accounted for by the predictors, species interactions and correlations driven by missing covariates. We show how estimation and inference for the considered models can be performed efficiently using either the Laplace or the variational approximation method. We use simulations to study the finite-sample properties of the two approaches. Examples are used to illustrate the methods.

E1137: **Making more out of ecological community data: A conceptual framework and its implementation as models and software**
Presenter: **Gleb Tikhonov**, University of Helsinki, Finland
Co-authors: Otso Ovaskainen
Community ecology aims to understand what factors determine the assembly and dynamics of species assemblages at different spatiotemporal scales. To facilitate the integration between conceptual and statistical approaches in community ecology, we propose Hierarchical Modelling of Species Communities (HMSC) as a general, flexible framework for modern analysis of community data. While non-manipulative data allow for only correlative and not causal inference, our framework facilitates the formulation of data-driven hypotheses regarding the processes that structure communities. We model environmental filtering by variation and covariation in the responses of individual species to characteristics of their environment, with potential contingency on species traits and phylogenetic relationships. We capture biotic assembly structure by species-to-species association matrices, which may be estimated at multiple spatial or temporal scales. We operationalize the HMSC as a Bayesian joint species distribution model, melding hierarchical regression part that accounts for fixed effects, latent factor components, capturing species associations, and flexible generalized modelling techniques. We present an efficient full conditional Gibbs block sampler for full Bayesian estimation of model parameters and implement the framework as R- and Matlab-packages. Armed with this tool, community ecologists can make sense of many types of data, including spatially explicit data and time-series data.

E0732: **Generalized linear latent variable models for the analysis of cognitive functioning over time**
Presenter: **Silvia Cagnone**, University of Bologna, Italy
Co-authors: Silvia Bianconcini
Dimensions of cognitive functioning are potentially important, but often neglected determinants of the central economic outcomes that shape overall well-being over the life course. The Health and Retirement Study and the Asset and Health Dynamic study (HRS/AHEAD) aim to examine the impact of cognitive performance and decline on key domains of interest(e.g., health and daily functioning, retirement, economic and health decision making). The analysis of the HRS/AHEAD cognitive data is performed using latent variable models that easily allow to determine common factors of the cognitive items and examine their dynamic over time. The estimation of these models is cumbersome when the observed cognitive items are of different nature as in the HRS/AHEAD study. Indeed, problems related to the integration of the likelihood function arise since analytical solutions do not exist. This problem is more evident in presence of longitudinal data when the number of latent variables increases proportionally to the number of items. We analyze the performance of a new integration method, known as Dimension Reduction Method (DRM), in the estimation of the latent individual cognitive status over time of the HRS/AHEAD data. It provides parameter estimates as accurate as techniques commonly applied in the literature, but without sharing the same computational complexity of the latter. We show that it can be applied in situations in which standard techniques are unfeasible.

E0901: **Item response theory model fit assessment via posterior predictive checking: Two case studies**
Presenter: **Mariagiulia Matteucci**, University of Bologna, Italy
Co-authors: Stefania Mignani
Within the framework of latent variable models, item response theory (IRT) models are used to analyze educational and psychological response data. The issue of model fit assessment is crucial especially in detecting violations to the assumption of unidimensionality and the consequent need of a multidimensional solution, where the conditional independence holds given a set of latent variables. To this aim, Bayesian posterior predictive checks may be employed when the model parameters are estimated through a Markov chain Monte Carlo algorithm. The results of a simulation study using discrepancy measures based on association or correlation among item pairs are presented. Moreover, the effectiveness of the method is shown in two different empirical applications. In the first case study, the perceived benefits and costs of residents towards the tourism industry are investigated by using a questionnaire under the assumption of a multidimensional latent structure. In the second case study, response data coming from an educational language assessment are taken into account where the examinees are scored on a single scale, even if the tests

explicitly consists of two different subtests.

EO744 Room MAL 540 OPTIMAL EXPERIMENTAL DESIGN FOR BIG DATA PROBLEMS Chair: HaiYing Wang

E1566: **Optimal design of experiments for networked data**
Presenter: **Ben Parker**, University of Southampton, United Kingdom
The problem of how to optimally design experiments for a large number of participants who are connected by a network structure is considered. As a motivating example, consider determining which of a selection of adverts on Facebook is effective, given that users have some friendship structure, as well as potentially some other characteristics which can be expressed as a blocking structure. We introduced previously a linear network effects model, and found optimal designs when experimental units were connected according to some relationship, which was specified by an adjacency matrix. We showed how networks could be useful in a variety of applications: for example, agricultural experiments where experimental units (plots) were connected by some spatial relationship, and also in crossover trials, where experimental units were connected by temporal networks. We develop now faster algorithms that allow optimal designs on networks to be found more quickly, and makes the problem of finding optimal designs for large networks manageable. We also argue that there is a wide class of experiments that can be reformulated into a problem of design on a network. By regarding experimental design as a problem in network science, we can improve experimental design algorithms for large networks, and also find designs even when there is no obvious network relationship.

E1548: **Optimal design theory: A device to select a good sample from big data**
Presenter: **Chiara Tommasi**, University of Milan, Italy
Co-authors: Laura Deldossi
Big Data are generally huge quantities of digital information accrued automatically and/or merged from several sources and rarely result from properly planned population surveys. A Big Dataset is herein conceived as a collection of information concerning a finite population. Since the analysis of an entire Big Dataset can require enormous computational effort, we suggest selecting a sample of observations and using this sampling information to achieve the inferential goal. Instead of the design-based survey sampling approach (which relates to the estimation of summary finite population measures, such as means, totals, proportions...) we consider the model-based sampling approach, which involves inference about parameters of a super-population model. This model is assumed to have generated the finite population values, i.e. the Big Dataset. Given a super-population model we can apply the theory of optimal design to draw a sample from the Big Dataset which contains the majority of information about the unknown parameters of interest.

E1556: **Orthogonalizing EM: A design-based least squares algorithm**
Presenter: **Shifeng Xiong**, Chinese Academy of Sciences, China
The aim is to introduce an efficient iterative algorithm, intended for various least squares problems, based on a design of experiments perspective. The algorithm, called orthogonalizing EM (OEM), works for ordinary least squares (OLS) and can be easily extended to penalized least squares. The main idea of the procedure is to orthogonalize a design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. We establish several attractive theoretical properties concerning OEM. For the OLS with a singular regression matrix, an OEM sequence converges to the Moore-Penrose generalized inverse-based least squares estimator. For ordinary and penalized least squares with various penalties, it converges to a point having grouping coherence for fully aliased regression matrices. Convergence and the convergence rate of the algorithm are examined. Finally, we demonstrate that OEM is highly efficient for large-scale least squares and penalized least squares problems, and is considerably faster than competing methods when n is much larger than p. Some follow-up work is also discussed.

EO597 Room MAL 541 NEW DEVELOPMENTS IN TIME SERIES ANALYSIS Chair: Stanislav Volgushev

E0399: **Bootstrap seasonal unit root test under seasonal heterogeneity**
Presenter: **Nan Zou**, University of Toronto, Canada
Co-authors: Dimitris Politis
Both seasonal unit root and seasonal heterogeneity are common in seasonal data. When testing seasonal unit roots under seasonal heterogeneity, it is unknown if we can apply existing tests designed for seasonal homogeneous settings, for example, HEGY test, and it is unclear what test we should implement if they fail. To answer these questions, the validity of augmented HEGY test and unaugmented HEGY test under seasonal heterogeneity is firstly analyzed. It turns out that the asymptotic null distributions of the HEGY statistics testing the single roots at 1 or -1 are standard and pivotable and are identical to the asymptotic null distributions under seasonal homogeneity. On the other hand, the asymptotic null distributions of the statistics testing any coexistence of roots are non-standard and not directly pivotable and are different from the asymptotic null distributions under seasonal homogeneity. Therefore, HEGY tests are not directly applicable to the joint tests for any concurrence of seasonal unit roots under seasonal heterogeneity. Bootstrap is secondly proposed as a remedy. In particular, we bootstrap augmented HEGY test with seasonal iid bootstrap and unaugmented HEGY test with seasonal block bootstrap. The consistency of these bootstrap procedures is established. The finite-sample behavior of these bootstrap tests is more desirable than their competitors'. These bootstrap tests are used to detect the seasonal unit roots in UK seasonal consumption data.

E1037: **Sequential detection of trend changes in irregularly observed panel data**
Presenter: **Tobias Kley**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz
Online surveillance of time series is traditionally done with the aim to identify changes in the marginal distribution under the assumption that the data between change-points is stationary and that new data is observed at constant frequency. In many situations of interest to data analysts, the classical approach is therefore too restrictive to be used unmodified. We propose a unified system for the monitoring of structural changes in streams of data where we use generalised likelihood ratio-type statistics in the sequential testing problem, obtaining the flexibility to account for the various types of changes that are practically relevant (such as, for example, changes in the trend of the mean). Our method is applicable to individual or large numbers of sequences where new observations are allowed to arrive irregularly and asynchronously across sequences. In an empirical illustration we apply the procedure to intra-day prices of components of the NASDAQ-100 stock market index.

E1343: **Gaussian approximation for high dimensional vector under physical dependence**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Guang Cheng
A Gaussian approximation result is developed for the maximum of a sum of weakly dependent vectors, where the data dimension is allowed to be exponentially larger than the sample size. Our result is established under the physical/functional dependence framework. This is as a substantive extension of previous work to time series based on a variant of Steins method developed therein.

E1485: **Two novel resampling strategies for dependent data**
Presenter: **Srijan Sengupta**, Virginia Tech, United States
Co-authors: Xiaofeng Shao, Stanislav Volgushev
Two novel resampling strategies are proposed, namely, the dependent random weighting (DRW) and the subsampled double bootstrap (SDB). The DRW is a generalization of the traditional random weighting where the weights are made to be temporally or spatially dependent and are adaptive to the configuration of the data. Unlike the block-based bootstrap or subsampling methods, the DRW can be used for irregularly spaced time series and spatial data without any implementational difficulty. The SDB is a fast resampling strategy for massive data applications which are increasingly prevalent. For massive datasets, classical bootstrap strategies (and its block-based versions) becomes prohibitively costly in computation even with modern parallel computing platforms. Recently a method called BLB (Bag of Little Bootstraps) for massive data has been proposed which is more computationally scalable with little sacrifice of statistical accuracy. Building on BLB and the idea of fast double bootstrap, we propose the SDB for both independent data and time series data. For both new methods, we establish theoretical properties and demonstrate empirical performance in simulation studies and data analysis.

EC693 Room MAL 153 CONTRIBUTIONS IN STATISTICAL MODELLING II Chair: George Tzougas

E1678: **Bivariate power- and envelope distributions originating from the elliptical class**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Andriette Bekker
A bivariate power distribution and its linked bivariate envelope distribution, originating from the elliptical distribution, is proposed. The marginals of the bivariate power distribution are of gamma type. Characteristics of these bivariate distributions, in particular, the density function, distribution function, and product moments are of special interest. The flexibility of this contribution is illustrated with the performance metric outage probability assuming different members of the underlying elliptical class, where the proposed bivariate power distribution may act as an alternative to existing fading models.

E1554: **The negative binomial inverse Gaussian regression model with an application to insurance ratemaking**
Presenter: **George Tzougas**, London School of Economics and Political Science, United Kingdom
Co-authors: Jun Ming Lim, Wei Li Hoon
The aim is to propose the Negative Binomial-Inverse Gaussian (NBIG) regression model as a competitive alternative to mixed Poisson regression models that have been widely used for actuarial purposes. The Negative Binomial-Inverse Gaussian regression model can be considered as a candidate model for highly dispersed count data and, to the best of our knowledge, this is the first time that it is used in a statistical or an actuarial setting. Specifically, the main contribution is that it illustrates that Maximum Likelihood (ML) estimation of the Negative Binomial-Inverse Gaussian regression model can be accomplished rather easily via an Expectation Maximization (EM) type algorithm. Moreover, the EM scheme we propose can be employed to estimate other members of the mixed Negative Binomial family as it can address situations where the mixing distribution is not conjugate to the Negative Binomial distribution. Furthermore, the a prori and a posteriori, or Bonus-Malus, premium rates resulting from the NBIG model are compared to those determined by the Negative Binomial Type I (NBI) and the Poisson-Inverse Gaussian (PIG) regression models that have been traditionally used for a priori and a posteriori ratemaking.

E1331: **A new family of estimators for the population mean using exponential functions in simple random sampling**
Presenter: **Ceren Unal**, Hacettepe University, Turkey
Co-authors: Cem Kadilar
In sampling theory, several ratio-type, product-type and regression-type estimators are widely used for the estimation of the mean. Alternatively, exponential estimators have also been proposed in recent years. The purpose is to propose more efficient estimators than existing estimators. For that, a family of estimators based on the exponential function is proposed for Simple Random Sampling. The expressions for the bias and Mean Square Error (MSE) of the proposed family are derived to the first degree of approximation. In addition, comparisons are made with some of the existing estimators. We prove that the proposed estimators are more efficient than existing estimators under the obtained conditions. Moreover, a numerical study is conducted to support theoretical results.

E0521: **Exploring the possibility of non-parametric estimation in the first stage of IV2SLS estimation**
Presenter: **Aslam Muhammad**, Bahauddin Zakariya University, Pakistan
Co-authors: Sajjad Haider Bhatti
The potential bias arising from the endogeneity is common in many empirical econometric relationships. The problem of endogeneity is largely tackled by Instrumental Variables Two Stage Least Squares (IV2SLS) estimation. The idea of estimating the first stage of IV2SLS non-parametrically is presented. The Mincer wage model based on human capital theory is widely used and is typically estimated by using instrumental variables 2SLS to address the possible bias due to endogenous schooling variable. We estimate the Mincer wage function using the French labour force data with recently proposed instrumental variable. We estimate the first stage schooling equation using LOESS method. Our findings show that the estimates from non-parametric first stage estimation in IV2SLS are more efficient compared to those from the traditional IV2LSL approach. Hence, the use of IV2SLS approach with non-parametric first stage is recommended for empirical studies dealing the issue of endogeneity bias.

EG010 Room MAL 151 CONTRIBUTIONS IN COPULAS Chair: Enrique de Amo

E1466: **Evaluating the power and significance levels of tests of copula symmetry**
Presenter: **Alphonce Bere**, University of Venda, South Africa
The power and significance levels of tests of copula symmetry/exchangeability are evaluated through a simulation study using copula models not previously considered. Khoudraji's device is used to obtain asymmetric versions of the copulas which are used in the power study. As in previous studies, it is shown that, for small sample sizes, the tests do not hold their nominal levels, that the Cramer-von Misus type statistics are the most powerful and that the power increases with sample size and degree of asymmetry in the data. Contrary to previous findings, our results do not reveal any hierarchy in terms of the power of the Jasson-type tests.

E0202: **Predictive inference for bivariate data using nonparametric copula**
Presenter: **Tahani Coolen-Maturi**, Durham University, United Kingdom
A new method is presented for prediction of an event involving a future bivariate observation. The method combines nonparametric predictive inference (NPI) applied to the marginals with a nonparametric copula to model and estimate the dependence structure between two random quantities, as such the method is fully nonparametric. In NPI, uncertainty is quantified through imprecise probabilities. Several novel aspects of statistical inference are presented. First, the link between NPI and copulas is powerful and attractive with regard to computation. Secondly, statistical methods using imprecise probability have gained substantial attention in recent years, where typically imprecision is used on aspects for which less information is available. A different approach, namely imprecision mainly being introduced on the marginals, is presented for which there is typically quite sufficient information, in order to provide robustness for the harder part of the inference, namely the dependence structure.

Thirdly, the set-up of the simulations to evaluate the performance of the proposed method is novel, key to these are frequentist comparisons of the success proportion of predictions with the corresponding data-based lower and upper predictive inferences. All these novel ideas can be applied far more generally to other inferences and models.

E1364: **Quasi copulas and the best possible bounds for a given set of copulas**
Presenter: **Enrique de Amo**, University of Almeria, Spain
Co-authors: Juan Fernandez Sanchez, Manuel Ubeda Flores
The best possible bounds on the set of copulas with a given value of the Spearman's footrule are studied. Unlike the case of bounds on sets of copulas with given value of other measures, these bounds are not necessarily copulas, but proper quasi copulas.

E1746: **Bayesian structured additive distribution regression with non-random sample selection**
Presenter: **Paul Wiemann**, University of Goettingen, Germany
Neglecting the selection process may lead to erroneous results, e.g. biased estimates, when non-randomly selected data is analyzed. Sample selection models account for the selection process and attempt to correct the selection bias by assuming a hierarchical procedure, where the first level governs the availability of observations in the second level. Well-established sample selection models make strong distributional assumptions regarding the dependency structure between both levels and furthermore, regarding the marginal distribution of the outcome of interest. Both constraints make them unsuitable for many practical applications. The presented approach addresses the issues of sample selection as well as both limitations. A copula is employed to entangle the selection process and the outcome of interest into a multivariate distribution. Since the marginal distributions are separated from the dependency structure, Bayesian distributional regression can be used to model the marginal distribution of outcome of interest appropriately. In the resulting joint model, structured additive predictors describe each parameter and thus allow for various effect types. The proposed Bayesian inference scheme uses an efficient Markov chain Monte Carlo technique to estimate the posterior distribution. An application from psychological research, which motivated the development of this model, serves as an illustrative example.

EG718 Room MAL 152 CONTRIBUTIONS IN CLUSTERING Chair: Christian Hennig

E1529: **Clustering aggregated data: The use of distances on distribution laws**
Presenter: **Margarida G M S Cardoso**, ISCTE-IUL, Portugal
Big data can be reduced into aggregated data using common summary statistics. The use of clustering procedures for aggregated data should, naturally, rely on adequate distances to compute heterogeneity between aggregated observations, e.g. considering histogram data. In this setting, distances on distribution laws can be particularly useful, although little work has been done in this area. A clustering analysis is conducted illustrating the use of three distances based on distribution laws on aggregated data. The data set considered originates from the European Social Survey and regards human values across regions in Europe. A K-Medoids algorithm is used. The results obtained are compared based on several indicators: e.g. within-between clusters average distance, average silhouette width, Calinski and Harabasz index and Dunn index. A congruent evaluation is conducted using the same distances both to build the clusters and evaluate them. In addition, indicators are also computed based on the commonly used Euclidean distance. Discussion refers to the choice a particular clustering solution.

E1717: **Multi-resolution clustering of time dependentfunctional data with applications to climatereconstruction**
Presenter: **Johan Strandberg**, Umea University, Sweden
Co-authors: Konrad Abramowicz, Lina Schelin, Sara Sjostedt-de Luna
A multi-resolution approach used to cluster dependent functional data is presented. Given a lattice of (time) points, a function is observed at each grid point. We assume that there are latent (unobservable) groups that vary slowly over time. We consider the case when at different time scales (resolutions) different groupings arise, with groups being characterised by distinct frequencies of the observed functions. We propose and discuss a non-parametric double clustering based method, which identifies latent groups at different scales. We present an application of the introduced methodology to varved lake sediment data, aiming at reconstructing winter climatic regimes in northern Sweden at different resolutions during the last sixthousand years.

E1715: **Visualization of clustering on multiple data**
Presenter: **Mariko Takagishi**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa
Visualization of clustering results for multiple data will be considered. We often obtain multiple data for the same items. By applying cluster analysis for each data, we can discover a latent structure in data. However, the relationship of clusters between different data cannot be interpreted if the cluster analysis is applied separately on each data. Therefore, a method to visualize the relationship between clusters obtained from different data is proposed. By using this method, clustering structure on a whole data can be visually understood.

E1425: **A clustering procedure for studying financial integration with big data time series**
Presenter: **Nuno Crato**, EC-JRC, Italy
Co-authors: Pilar Poncela, Jorge Caiado
Time and frequency-domain procedures for characterizing and comparing large sets of long time series are studied. The procedures are computationally simple, condense second-order information of the time series under consideration, and develop similarity and dissimilarity measures for comparing these time series. On the basis of these measures, we exemplify and compare various ways of clustering time series. Instead of using all the information available from data, which would be computationally very expensive, we propose some regularization rules in order to select and summarize the most relevant information for clustering purposes. We use these procedures to study the evolution of several stock markets indices, extracting information on the European financial integration. We further show the effect of the recent financial crisis over these indices behavior.

CO467 Room Chancellor's Hall MACHINE LEARNING TECHNIQUES FOR TIME SERIES FORECASTING Chair: Lyudmila Grigoryeva

C1512: **Adaptive online model selection for linear and non-linear AR**
Presenter: **Stephane Chretien**, NPL, United Kingdom
Model selection is an important task which may have a tremendous impact on time series forecasting. Unfortunately, model selection can be computationaly expensive, especially in the case of non-linear models such as models governed by complex deep networks. We present a new approach to model selection using the Hedge algorithm. We will analyse the theoretical performance of the method in the linear case and in some non-linear cases. We will show computational experiments that demonstrate the effectiveness of the proposed approach in various settings. In particular, experimental results using a recent approach will be presented.

C1823: **Neural networks for asynchronous time series**
Presenter: **Mikolaj Binkowski**, Imperial College London and Hellebore Capital Ltd., United Kingdom
Many real-world time series are asynchronous in the sense that their separate dimensions are observed at different and irregular moments of time. However, most of approaches to discrete time series or continuous stochastic processes require at least synchronous observations. As a solution to

this problem, we propose Significance-Offset Convolutional Neural Network, a deep convolutional network architecture inspired by autoregressive models and gating mechanisms used in recurrent neural networks. The architecture uses a weighting system designed for asynchronous inputs, where the final predictor is obtained as a weighted sum of adjusted regressors, while the weights are data-dependent functions learnt through a convolutional network. We evaluate it on a hedge fund proprietary dataset of over 2 million quotes for a credit derivative index, an artificially generated noisy autoregressive series and household electricity consumption dataset. The proposed architecture achieves promising results as compared to convolutional and recurrent neural networks.

C1594: **Predicting emerging market sovereign credit spreads with machine learning/data science techniques**
Presenter: **Gary Anderson**, Federal Reserve Board, United States
Co-authors: Alena Audzeyeva
Standard financial time series regression analysis has previously been applied to construct a collection of linear models for forecasting credit spreads of sovereign debt issued by Brazil, Mexico, Philippines and Turkey for the period just before and just after the Lehman crisis. We apply a variety of Support Vector Machine Regression (SVMR) kernels to develop a family of parsimonious non-linear regression models. We estimate the model parameters using data science motivated cross-validation techniques as well as with traditional non-linear estimation techniques. We find that these models can significantly outperform the linear models in reality check tests and tests for superior predictive ability (SPA).

C1572: **Welfare maximizing dynamic treatment allocation and recommendation**
Presenter: **Bezirgen Veliyev**, Aarhus University, Denmark
Co-authors: Anders Kock, David Preinerstorfer
In many decision problems information arrives gradually. For example, patients with a certain illness arrive gradually to the hospital and information about competing treatments accumulates as patients are treated. The hospital thus faces a tradeoff between exploring the merits of the available treatments and administering the optimal treatment as often as possible. We cast this problem as a multi-armed bandit problem and develop treatment algorithms that are optimal in the sense that no other algorithm can incur a smaller loss in terms of rates. Furthermore, we show that by the end of the treatment period we can give a recommendation on the best treatment which is not much worse than if this treatment had been known. We take into account that the optimal treatment may be person specific and that one may not only target the treatment with the highest expected effect as its risk may also be of interest. More precisely, we show minimax optimality of our treatment algorithm even when targeting quite general functionals of the distribution of treatment outcomes. These functionals include quantiles and smooth functions of a finite number of moments of the treatment outcomes.

CO732 Room Court TIME SERIES COPULA MODELLING Chair: Michael Smith

C1286: **Predicting US federal debt limit crises and the policy responses**
Presenter: **Anthony Garratt**, University of Warwick, United Kingdom
Co-authors: Shaun Vahey, Liz Wakerly
The deployment of extraordinary measures during the summer of 2011 allowed the US Federal government to function, temporarily circumventing the debt ceiling. This crisis generated considerable press attention, sparking debate about the impacts of a Federal government shut-down, and was partially resolved with the passing of the Budget Control Act of 2011 in August. We compute the probability of exceeding the debt ceiling using a Gaussian copula model with non-parametric marginals. The asymmetric predictive densities are derived from two monthly forecasting specifications: a copula model with time-invariant autoregressive dependence; and, a copula model with time-varying autoregressive dependence using rolling windows. We show that the variant with rolling windows comfortably outperforms both a rolling window Gaussian autoregressive model and the expanding window copula model in terms of out of sample density forecasting performance. Furthermore, our approach provides the probabilistic information necessary to predict the policy actions taken by the US Treasury to limit the economic impact of the debt limit crisis.

C1302: **Inversion copulas for realized GARCH models**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Inversion copulas show promise in modelling latent nonlinear state space models with Markov dependence structures. We extend this idea to cover nonlinear time series with non-Markov dependence, with focus on two special cases: the well-known GARCH and Realized GARCH specifications. Both present challenges in finding and evaluating the implied margin of the latent variable: we discuss some possible solutions here. Likelihood and Bayesian computational methods are derived for estimation, inference and forecasting purposes. The sampling properties of these estimators are illustrated via a simulation study. The two new time series inversion copula models are used to model and forecast financial returns from several financial indices, including an emerging markets index and a gold and silver index. The proposed models are competitive for density and tail risk forecasting in these series, compared to a range of popular, competing financial time series models.

C1301: **Bernanke vs Taylor: US monetary policy rules with non-Gaussian marginal distributions**
Presenter: **Ozer Karagedikli**, Reserve Bank of New Zealand, New Zealand
Co-authors: Shaun Vahey, Liz Wakerly
Although John Taylor has argued that his policy rule provides a useful a benchmark for monetary policy, Ben Bernanke and others have deployed many variants of the Taylor rule. Using Taylor's preferred specification, U.S. monetary policy was too loose 2003 to 2005, contributing to the build up to the Great Recession. In contrast, given Bernanke's "best" model, nominal interest rates through the same period look about right. The aim is to re-examine the path of nominal interest rates through the Great Recession and the aftermath, reformulating the Taylor rule to accommodate the non-Gaussian marginal distribution of nominal interest rates. Assuming linear dependence between the macroeconomic variables of interest, we find strong evidence to support Bernanke's account of monetary policy leading up to the recession.

C1282: **Real-time macroeconomic forecasting with a heteroskedastic inversion copula**
Presenter: **Michael Smith**, University of Melbourne, Australia
Accounting for asymmetries in the forecast densities of macroeconomic variables can improve their accuracy. In multivariate time series, this can be achieved by using a copula to capture both serial and cross-sectional dependence, allowing the margins to be modeled directly as nonparametric. Yet most existing high-dimensional copulas cannot capture heteroskedasticity. To do so, we propose a new copula created by the inversion of a multivariate unobserved component stochastic volatility model, and show how to estimate it using Bayesian methods. We study its real-time forecasts and their accuracy for four quarterly U.S. macroeconomic variables. The copula model captures heteroskedasticity, dependence in the level, time-varying asymmetry and heavy tails, bounds on the variables and other features. Over the window 1975Q1 – 2016Q3, the point and density forecasts out-perform those from a range of benchmark models, particularly for the nowcast of GDP growth.

CO194 Room G11 CYCLICAL PROPERTIES OF FINANCIAL AND ECONOMIC DATA Chair: Jozef Barunik

C0809: **Low-frequency macroeconomic risks and asset prices: A critical appraisal of Epstein-Zin preferences**
Presenter: **Georgios Xyngis**, Cardiff Business School, United Kingdom
The aim is to examine if low-frequency macroeconomic growth and volatility risks are priced in asset prices. The motivation comes from the spectral decomposition of the pricing kernel under recursive preferences. We demonstrate that macroeconomic shocks with frequencies lower than the business-cycle are not robustly priced in the cross-section of expected returns. In addition, the estimated risk premia are economically small, have wrong signs and the low-frequency risk exposures fail to match known patters in average returns. Overall, we highlight the need for risk preferences that allow investors to be more risk averse to business-cycle frequencies.

C0424: **On the importance of supply shocks for real exchange rates: A fresh view from the frequency-domain**
Presenter: **Britta Gehrke**, University of Erlangen-Nuremberg, Germany
Co-authors: Fang Yao
The role of supply shocks for real exchange rate fluctuations is re-examined and insights from the frequency domain perspective are exploited. In contrast to the existing literature, we find that supply shocks are one important driver of US real effective exchange rate fluctuations at low frequencies, while real demand shocks matter the most at high and medium frequencies. In addition, we propose an approach to structurally decompose the persistence of the real exchange rate and find that supply shocks explain up to half of its persistence.

C0511: **On simple sum monetary statistics**
Presenter: **Michael Ellington**, University of Liverpool, United Kingdom
The aim is to investigate how defective simple sum monetary aggregates can be. In doing so, we compare Divisia monetary statistics, which are measured using a micro-theoretic approach, with their simple sum counterparts. We provide a comprehensive reduced-form and structural analysis of time-varying parameter VAR models of the US economy using a range of monetary aggregates. Reduced-form results show a strong link between Divisia money and macroeconomic fundamentals at low and medium frequencies, whilst also providing substantial evidence supporting the conjecture that forecasters, and indeed central banks, should replace simple sum measures of money with their Divisia counterparts. Structural analysis uncovers economically significant and statistical differences in the structural impact coefficients of money within the monetary policy rule. Frequency domain variance decompositions report substantial contributions of monetary policy shocks to real GDP growth and inflation, both at an infinite horizon, and business cycle frequency. Although there are economic differences in the contribution these shocks conditional on the monetary aggregate used, we find no statistical differences. For policymakers, these results can help guide monetary policy reactions to recessions with an idea of the implications these shocks have on macroeconomic variation at different frequencies.

C0977: **Frequency response analysis of monetary policy transmission**
Presenter: **Lubos Hanus**, Charles University, Czech Republic
Co-authors: Lukas Vacha
A new approach is considered to look at the effects of economic shocks to dynamics of economic systems. We analyse the widely known phenomenon of price puzzle in a time-varying environment using the frequency decomposition. We use the frequency response function to measure the power of a shock transferred to different economic cycles. Considering both the time-variation of the system and frequency analysis, we can quantify the local dynamics of shocks at given time and over frequencies and thus reveal policy implications the system can provide. While studying the monetary policy transmission of the U.S., the empirical evidence shows that low-frequency cycles are prevalent, however, their amplitudes vary significantly in time.

CO079 Room G5 NEW DEVELOPMENTS IN MACRO AND FINANCIAL ECONOMETRICS Chair: Claudio Morana

C1816: **Time varying persistence in GARCH-in-mean models with time-dependent coefficients**
Presenter: **Alessandra Canepa**, Brunel University, United Kingdom
A number of unit root tests are considered and the results of a Monte Carlo experiment are presented to investigate the size and power properties of these tests in the presence of breaks in the mean and the variance equation of an AR-GARCHM data generating process. It is found that the location and the magnitude of the breaks badly affect the size and power properties of the test statistics. In particular, in the presence of an in-mean effect, conventional unit root tests tend to falsely indicate that the underlying process is I(1).

C1400: **Financial crises and the dynamic linkages between stock and bond returns**
Presenter: **Faek Menla Ali**, University of Sussex, United Kingdom
The aim is to investigate the dynamic linkages in terms of the first and second moments between stock and bond returns, within a wide range of advanced economies, over the different phases of the recent financial crisis. The adopted empirical framework is a bivariate volatility model, where volatility spillovers of either positive or negative sign are allowed for. Our results lend support to the existence of a substantial time-variation in the dynamic linkages between these financial assets over the different stages of the recent crisis. Moreover, they also point to a remarkable variety in spillovers between stock and bond markets across countries. In particular, our results of the return spillovers show that such spillovers mostly run from stocks to bonds and exhibit a time-varying pattern over all three stages of the crisis in most countries. Regarding the volatility spillovers, such spillovers from bond returns to those of stocks are stronger than the other way round and also exhibit a time-varying pattern in most countries. Furthermore, the portfolio performance comparison results show that by considering time-varying return and volatility spillovers when calculating the risk-minimising portfolio weights of the selected assets, the portfolio volatility can be reduced despite limited diversification opportunities within national markets in times of financial crises.

C1435: **On a number of time series econometric issues of some importance**
Presenter: **Menelaos Karanasos**, Brunel University, United Kingdom
Co-authors: Alexandros Paraskevopoulos
A new innovative methodology for analysing time series models without the need to work with lag polynomials is presented. Following laborious research work, the literature contains a diversity of linear time varying models (i.e., smooth transition AR specifications, GARCH processes with time dependent coefficients, Markov switching models, generalized random coefficients AR processes, periodical and cyclical formulations) whose time series properties remain unexplored. Making progress in interpreting seemingly different models requires us to provide a common platform for the investigation of their time series properties. We develop a theoretical foundation on which work in synthesizing these formulations can be done. With the help of a few detailed examples, i.e., GARCH in mean models with abrupt breaks, we demonstrate how to encompass various time series models within our unifying theory. Our innovative methodology allows us to study stochastic linear difference equations of ascending order and handle time varying specifications of infinite order. An advantage of our methodology is that it can be applied with ease in a multivariate setting and provide a solution to the problem at hand without adding complexity. The main strength of our general solution and the way we have expressed it, is that researchers can use it for a multiplicity of problems. The significance of our methodology is almost self-evident from the large number of problems that can solve.

C0206: **Semiparametric estimation of multivariate GARCH models**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

A new simple semiparametric estimator of the conditional variance covariance and correlation matrix (SP-DCC) is introduced. While sharing a similar sequential approach to existing dynamic conditional correlation (DCC) methods, SP-DCC has the advantage of not requiring the direct parameterization of the conditional covariance or correlation processes, therefore also avoiding any assumption on their long-run target. In the proposed framework, conditional variances are estimated by univariate GARCH models, for actual and suitably transformed series, in the first step; the latter are then nonlinearly combined in the second step, according to basic properties of the covariance and correlation operator, to yield nonparametric estimates of the various conditional covariances and correlations. Moreover, in contrast to available DCC methods, SP-DCC allows for straightforward estimation also for the non-symultaneous case, i.e., for the estimation of conditional cross-covariances and correlations, displaced at any time horizon of interest. A simple ex-post procedure, to ensure well behaved conditional covariance and correlation matrices, grounded on nonlinear shrinkage, is finally proposed. Due to its sequential implementation and scant computational burden, SP-DCC is very simple to apply and suitable for the modeling of vast sets of conditionally heteroskedastic time series.

CO120 Room Jessel SPATIAL ECONOMETRIC MODELLING Chair: Maria Kyriacou

C0859: **Threshold regression with social interactions**
Presenter: **Andros Kourtellos**, University of Cyprus, Cyprus
Co-authors: Antri Konstantinidi, Yiguo Sun

A new class of social interaction models that generalize the spatial autoregressive model is developed in order to allow for threshold effects. Threshold-type nonlinearity is one of the most interesting forms of nonlinearity as it is often implied by economic theory. In particular, we study identification and develop a statistical theory for the threshold parameter as well as the regression coefficients. We exploit the non-linear structure of the threshold regression as well as the network structure to overcome identification issues that arise in the linear-in-means model and propose a GMM approach for the estimation. We assess the performance of our methods using a Monte Carlo simulation and provide an empirical illustration.

C1143: **Spatial heterogenous autoregression with varying-coefficient covariate effects**
Presenter: **Maria Kyriacou**, University of Southampton, United Kingdom
Co-authors: Zudi Lu, Peter CB Phillips

The traditional SARX models offer a simple way of capturing the essence of spatial interactions via the Wy operator, but have been subject to criticism owing to their several limitations, including its inability to capture spatial non-linearities and unobserved heterogeneity. We propose a spatial heterogenous autoregressive exogenous (SHARX) model captures for such non-linearities and unobserved heterogeneity by allowing for varying-coefficents in both the exogenous regressors coefficients and to the error term structure. The coefficients of the exogenous regressors are allowed to smoothly vary with location s (which s denotes the denotes the "smoothing-parameter") and therefore enables us to introduce spatialtrends/non-stationarity in y or heterogenous non-linearity between X and s. We allow both the exogenous regressors and the innovation sequence to depend on location s by defining them as unknown functions of this 2-dimensional vector. Following a set of assumptions, the unknown parameters ate estimated by a profile maximum likelihood which is based on a two-step procedure where: 1. The unknown parameters are estimated at location s by by local maximum likelihood estimation (LMLE) for a given lambda, and 2. The the spatial profile likelihood can be defined from (1.) and the estimator of the spatial parameter is then defined as the maximum profile likelihood estimator (MPLE).

C1164: **On a data-driven semiparametric nonlinear model with penalized spatio-temporal lag interactions**
Presenter: **Zudi Lu**, University of Southampton, United Kingdom
Co-authors: Dawlah Alsulami, Zhenyu Jiang, Jun Zhu

Studying a possibly nonlinear impact of consumer price index (CPI) on the housing price at a state level in the USA, ignoring or misspecifying the temporal lag effects of the housing price both from the own state and the neighboring states, will result in a biased modelling. We therefore propose a data-driven semiparametric nonlinear time series regression model that accounts for spatio-temporal lag interactions in both space and time. A semiparametric penalised estimation procedure is suggested by utilising adaptive lasso to estimate the most important spatio-temporal lag interactions. Theoretical justification for the estimation procedure is developed. Empirical application to a USA data set demonstrates that the proposed method can substantially improve the estimation of the nonlinear impact of the monthly increment of CPI on the housing price return at a state level, with interesting spatio-temporal lag interactions identified.

C1290: **Adjusted maximum likelihood inference for spatial panel data models**
Presenter: **Federico Martellosio**, University of Surrey, United Kingdom

In a likelihood framework, an often successful way to deal with incidental parameters is to "adjust" the profile score of the parameter of interest. We consider the adjusted quasi-maximum likelihood estimator (QMLE) of the spatial parameter in a spatial panel model with individual and/or time fixed effects. The adjusted QMLE coincides with the QMLE if covariates are not present in the model. When covariates are present, the adjusted QMLE can be more accurate than the QMLE. Saddlepoint confidence intervals for the spatial autoregressive parameter based on the adjusted QMLE are proposed. In simulation, they perform very well against other higher-order methods.

CO545 Room Montague FINANCIAL ECONOMETRICS Chair: Wei Wei

C0972: **A regime-switching stochastic volatility model for forecasting electricity prices**
Presenter: **Peter Exterkate**, University of Sydney, Australia
Co-authors: Oskar Knapik

In a recent review paper, several crucial challenges outstanding in the area of electricity price forecasting are pinpointed. The aim is to address all of them by (i) showing the importance of considering fundamental price drivers in modeling, (ii) developing new techniques for probabilistic (i.e. interval or density) forecasting of electricity prices, and (iii) introducing an universal technique for model comparison. We propose a new regime-switching stochastic volatility model with three regimes (negative jump or "drop", normal price or "base", positive jump or "spike") where the transition matrix depends on explanatory variables. Bayesian inference is employed in order to obtain predictive densities. The main focus is on short-term density forecasting in the Nord Pool intraday market. We show that the proposed model outperforms several benchmark models at this task.

C1299: **A noise-robust trade classification algorithm**
Presenter: **Simon Jurkatis**, Freie Universitaet Berlin, Germany

A new trade classification algorithm is proposed that provides accurate estimation of the trade initiator in data environments that have challenged the literature on trade classification. These environments are imprecise timestamps relative to the frequency of quote changes and misalignments of trade times and their corresponding quote changes. Using data from Nasdaq's limit order book, which provides information of the trade initiator, the method is compared against the common alternatives under various data environments by artificially decreasing timestamp precision and adding noise to recorded transaction times for more than 130m trades. The results show that the new algorithm is the dominant choice with improvements

of up to reducing misclassification by half. The empirical relevance of these differences is demonstrated by estimating various measures of liquidity.

C1218: **Market maker inventory, bid/ask spreads, and the computation of option implied risk measures**
Presenter: **Daniela Osterrieder**, Rutgers Business School, United States
Co-authors: Bjorn Eraker, Ivan Shaliastovich
Option implied risk measures (OIRMs) such as the VIX index are computed from the midpoint of best bids and asks. For these measurements to be unbiased, the midpoints are implicitly assumed to be unbiased estimates of a "true" underlying option price. We derive a model where market makers adjust their spreads according to the size of their inventory. The model implies that negative inventory leads to upwardly biased midpoints, highly volatile asking prices, and right skewed spreads. We rely on a generalized method of moments (GMM) estimation procedure to show that these properties are true in the data set, which is comprised of detailed option-market observations and high-frequency measures of the underlying stock market. We further show that the OIRM's are very sensitive to the spread. For example, the ability of the variance risk premium (VRP) and higher order option implied tail measures in predicting stock returns is very sensitive to whether the they are computed from the bid, ask or midpoint.

C0915: **Identifying uncertainties from multiple factors: A study on electricity price**
Presenter: **Wei Wei**, Monash University, Australia
Co-authors: Asger Lunde
Using a multi-factor model, the uncertainties in electricity spot prices are separated into three risk factors: spikes, base-signals, and stochastic volatility. The model is estimated using the particle Markov chain Monte Carlo method and is applied to the Germany/Austria electricity market. Our results indicate that spike shocks are large and infrequent, and they usually die out within a day, while base-signals are more persistent than spikes. Moreover, the observed clustering of large price movements is explained by stochastic volatility. We apply our estimates in the spot market to the futures market and find that different risk factors have distinct impacts on futures prices. In particular, the base-signal dynamics govern the futures price dynamics. Lastly, we find evidence that the market price of both spike risk and base-signal risk are time-varying.

CO312 Room Senate RECENT ADVANCE FOR NON-LINEAR TIME SERIES Chair: Weining Wang

C0302: **A new semiparametric estimation of large dynamic covariance matrix with multiple conditioning variables**
Presenter: **Jia Chen**, University of York, United Kingdom
Co-authors: Degui Li, Oliver Linton
Estimation of dynamic covariance matrix with multiple conditioning variables is studied, where the matrix size can be ultra large (divergent at an exponential rate of the sample size). We introduce an easy-to-implement semiparametric method to estimate each entry via model averaging marginal regression, and then apply the shrinkage idea to obtain the large dynamic covariance matrix estimation. Under some regularity conditions, we derive uniform consistency results for the proposed estimators with general convergence rates. Furthermore, we consider generalising our methodology to deal with the case when the dimension of conditioning variables is diverging, and discuss an application to the dynamic optimal portfolio choice. Some numerical studies including simulation and empirical application are given to examine the finite-sample performance of the developed methodology.

C1149: **SVARs with breaks: Identification and inference**
Presenter: **Toru Kitagawa**, University College London, United Kingdom
Co-authors: Emanuele Bacchiocchi
The aim is to study identification of structural vector autoregressions (SVARs) with structural breaks in structural error variances and/or structural coefficients. We first provide the point-identifying conditions for the structural parameters in the presence of structural breaks, which generalizes the existing results in the literature. We then investigate scenarios where the point-identifying assumptions are relaxed and the structural parameters are only partially identified. We provide analytical characterizations of the identified set and discuss how to compute them given knowledge of reduced-form parameters. A inferential analysis is proposed based on robust Bayesian techniques.

C0823: **Varying random coefficient models**
Presenter: **Christoph Breunig**, Humboldt-Universitat zu Berlin, Germany
A linear model with varying random coefficients (VRC) is considered. VRCs are modeled additively separable with an unknown nonlinear function of covariates and an unobservable part. The VRC model allows for heterogeneous marginal effects which might vary with covariates. Identification of the distribution of marginal effects is established. The estimator is based on weighted sieve minimum distance. Its L_2 rate of convergence is derived. Pointwise and uniform limit theory of functionals is derived. Our estimator is easy to implement and performs well in finite sample.

C0290: **Long-range dependent curve time series**
Presenter: **Degui Li**, University of York, United Kingdom
Co-authors: Peter Robinson, Han Lin Shang
Methods and theory for functional time series with long-range dependence are introduced. The temporal sum of the curve process is shown to be asymptotically normally distributed. We show that the conditions for this cover a functional version of fractionally integrated autoregressive moving averages. We also construct an estimate of the long-run covariance function, which we use, via functional principal component analysis, in estimating the orthonormal functions spanning the dominant sub-space of the curves. In a more general, semiparametric context, we propose an estimate of the memory parameter, and derive its consistency result. A Monte-Carlo study of finite-sample performance is included, along with two empirical applications. The first of these finds a degree of stability and persistence in intra-day stock returns. The second finds similarity in the extent of long memory in age-specific fertility rates across some developed countries.

CO655 Room SH349 DESIGN OF REAL-TIME FILTERS FOR ECONOMIC MONITORING AND FINANCIAL TRADING Chair: Marc Wildi

C0490: **Basel III and the prediction of financial crises**
Presenter: **Simon van Norden**, HEC Montreal, Canada
Co-authors: Marc Wildi
Basel III allows national regulators to adjust the minimum level of reserves held by financial institutions in response to changing perceptions of the fragility of the financial system. This requires regulators to forecast potential systemic financial crises. The extent to which this is feasible is the subject of considerable controversy, not least because of its international importance to current banking regulation. The aim is to contribute to this debate by examining the extent to which regulators' preferred variables can usefully forecast systemic banking crises.We make two original contributions. First, we study how aspects of existing real-time measures of credit cycles affect the predictability of systemic banking crises. We show that simple theory-based modifications to filtering measures recently mandated by regulators can have an important effect on predictive

power. We then study the performance of alternatives based on optimal filter designs. We discuss the results in light of the current debate over the expected performance of credit cycles as guide for Countercyclical Capital Buffers.

C0799: **Real-time signal extraction of vector time series via multivariate direct filter analysis**
Presenter: **Tucker McElroy**, Census Bureau, United States
Co-authors: Marc Wildi

Real-time signal extraction for multivariate time series attempts to utilize information from related series in order to extract trends, cycles, or other patterns of interest using only present and past observations. When signal content is obfuscated by noise in the series of interest, but is more salient in related time series, a substantial improvement in extraction results can be expected. Multivariate direct filter analysis (MDFA) avoids using model-based concurrent filters, being predicated instead on using a frequency domain characterization of multivariate time series. We explicitly show how nonstationary effects, level constraints, and time shift constraints can be accounted for in the MDFA filter. While model-based frameworks can be replicated by the MDFA methodology, more complicated structures (e.g., nonlinear models) can also be entertained, thereby illustrating the flexibility of these nonparametric techniques. We demonstrate the power of the new methods on construction and employment data.

C0567: **Momentum and trend-following: Combining academia and industry**
Presenter: **Joerg Osterrieder**, ZHAW, Switzerland

Momentum trading strategies are thoroughly described in the academic literature and used in many trading strategies by hedge funds, asset managers, and proprietary traders. The hedge fund Man/AHL has made its internal strategy available and describes a momentum strategy for different asset classes in great detail from a practitioners point of view. Using a geometric Brownian Motion for the dynamics of the returns of financial instruments, we extensively explain the motivation and background behind each step of a momentum trading strategy. Constants and parameters that are used for the practical implementation are derived in a theoretical setting and deviations from those used by Man/ AHL are shown. The trading signal is computed as a mixture of exponential moving averages with different time horizons. We give a statistical justification for the optimal selection of the time horizons. Furthermore, we test our approach on global currency markets, including G10 currencies, emerging market currencies, and cryptocurrencies. Both a time series portfolio and a cross-sectional portfolio are considered. We find that the strategy works best for traditional fiat currencies when considering a time series based momentum strategy. This is the first comprehensive study showing both the underlying statistical reasons of how such trading strategies are constructed in the industry as well as empirical results using a large universe of currencies, including cryptocurrencies.

C0594: **Customized signal extraction: An application to FX-trading**
Presenter: **Marc Wildi**, Zurich University, Switzerland

Momentum strategies are extensively used in financial trading. They rely on the assumption that historical trends tend to live on and thus can be extrapolated into the near future. Formally, classic momentum filters can be interpreted in terms of optimal forecast rules in the context of particular model assumptions (about the data generating process of the series). We adopt a more general signal extraction perspective, whereby filters are the outcome of a real-time trend approximation problem. In this framework, particular filter properties which are felt to be relevant in financial trading, such as timeliness, smoothness or accuracy, can be emphasized explicitly. The approach is illustrated by an application to daily FX-data.

CG088 Room Bloomsbury CONTRIBUTIONS IN MULTIVARIATE MODELLING OF TIME SERIES Chair: Alain Hecq

C0401: **Time and the price impact of trades in Australian banking stocks around interest rate announcements**
Presenter: **Manh Cuong Pham**, Monash University, Australia
Co-authors: Heather Anderson, Huu Nhan Duong, Paul Lajbcygier

A nonlinear vector-autoregressive framework for trade durations, trade attributes (signs and volumes) and returns is proposed which incorporates the dynamic interdependence amongst these variables and relaxes the exogeneity assumption that is often imposed on durations in previous studies. We employ this framework to examine the role of durations and trade attributes in the price formation process for Australian banking stocks around interest rate announcements. We find that durations are not only correlated but also jointly determined with trade characteristics and returns. Shorter durations increase the price impact and autocorrelation of trades. Transactions executed within one minute around the announcements have shorter durations and larger impact on prices. Without a duration shock, the cumulative price impact of an unexpected trade is similar, regardless of whether or not durations are endogenously modeled. However, conditioning on an average before-announcement history, the cumulative price impact is higher (lower) following a negative (positive) duration shock if durations are endogenous, yet it stays unchanged if durations are treated as exogenous. Duration shocks contribute significantly less to the forecast error variance of returns than trade attributes shocks. The contribution of duration shocks is larger on announcement days, when durations are endogenously modeled, and for less liquid stocks.

C1755: **Low volatility of alternative UCITS: Fact or fiction**
Presenter: **Nabil Bouamara**, KU Leuven, Belgium
Co-authors: Kris Boudt, David Ardia

The (il)liquidity of alternative investments is attracting much attention from investors and regulators. However, from the viewpoint of fund structures this is a hard concept to define and measure. To overcome this problem, it has been previously establishes the classical connection between liquidity and autocorrelation for fund structures and showed that smoothed returns (due to pricing ambiguity) may give a deflated view of the true volatility of underlying returns. The model is based on a finite moving average model under a summation constraint and implicitly assumes bounded parameters. Within this framework, the maximum likelihood estimator is prone to estimation error. We provide a methodological and financial research contribution in terms of modelling the parameter structure and propose a structured estimation approach. Contrary to the standard unconditional model, we use commonality explained by fund characteristics and a nearest moment estimation. Finally, we illustrate the methods usefulness on a set of regulated liquid alternatives.

C1712: **Forecasting and policy analysis with trend-cycle bayesian VARs**
Presenter: **Jan Bruha**, Czech National Bank, Czech Republic
Co-authors: Michal Andrle

Trend-Cycle Bayesian VARs (TC-BVARs) are introduced for use in macroeconomic forecasting and policy analysis. Economic theory supports the view that trends and cycles are dominated by different shocks and transmission channels. Each variable is decomposed into trend and cyclical components. The flexibility of TC-BVARs comes from the fact that the model specifies flexible processes for low-frequency movements (trends) of variables and flexible VAR process for the cyclical frequencies. There is a clear distinction of cycles, trends, or exogenous time-varying policy targets. TC-BVARs benefit from the flexibility of VARs and from careful anchoring of the models long-run behavior. The state-space form of the model helps to work with missing data, mixed frequencies, and various forms of expert judgment and conditional forecasting. Structural TC-BVARs benefitfrom less biased reduced-form model specification.

C1774: **Wavelet variance ratio test and wavestrapping for the determination of the cointegration rank**
Presenter: **Burak Alparslan Eroglu**, Istanbul Bilgi University, Turkey

A wavelet based cointegration test for fractionally integrated time series is proposed. The test is non-parametric and asymptotically invariant to different forms of short run dynamics. The use of wavelets allows one to take advantage of the wavelet based bootstrapping method particularly known as wavestrapping. In this regard, we introduce a new wavestrapping algorithm for multivariate time series processes, specifically for cointegration tests. The Monte Carlo simulations indicate that this new wavestrapping procedure can alleviate the severe size distortions which are generally observed in cointegration tests with time series containing innovations with highly negative MA parameters. Further, we apply the the proposed methodology to analyze long run co-movements in the Credit Default Swap market of European Union Countries.

CG078 Room G4 CONTRIBUTIONS IN UNCERTAINTY AND FORECASTINGS IN CENTRAL BANK Chair: Svetlana Makarova

C0348: **On the rationality of expert's forecasts: An empirical insight from consensus economics data**
Presenter: **Frederique Bec**, THEMA University of Cergy-Pontoise and CREST, France
Co-authors: Raouf Boucekkine, Caroline Jardet

A theoretical model of forecasts formation is proposed which shows that the combination of observation costs and forecast changes costs can generate optimal forecast stickiness under rational expectations hypothesis. This model main predictions are that i) forecast updates are both time- and state-dependent, and ii) when the gap between the optimal and the actual forecast crosses a threshold, the forecast update is triggered immediately and it closes the gap. Monthly Consensus Economics survey data for French and German professional forecasters inflation forecasts provide evidence of time- and state-dependence of forecast updates as well as of the threshold-type dynamics of the latter. These data also support the co-existence of both types of costs with forecast change cost lower than observation cost.

C1418: **A suite of inflation forecasting models**
Presenter: **Luis J Alvarez**, Bank of Spain, Spain
Co-authors: Isabel Sanchez

A set of econometric models is described which is used to forecast consumer price inflation, a topic of renewed interest following the Great Recession and its aftermath. First, we consider highly disaggregated univariate time series models. Second, we propose transfer function models that include explanatory variables, such as oil prices or unit labour costs and that also allow for a slowly evolving local mean. Third, we consider macro-based models, including hybrid New Keynesian Phillips curves. These models have a theoretical foundation and include forward looking elements. Our main findings are that successful inflation forecasting models have to account for a slowly evolving local mean, in order to be able to cope with changes in trend inflation, as in the period after the Great Recession. Furthermore, differences in the features of product markets suggest that it is relevant to employ some sort of disaggregation to deal with heterogeneity in price setting. Finally, transfer function models tend to show a better forecasting performance than alternative models.

C1522: **Measuring the productivity and the efficiency of the Mexican Central Bank**
Presenter: **Rafael Bernardo Carmona Benitez**, Investigaciones y Estudios Superiores, S. C., Mexico
Co-authors: Maria Rosa Nieto Delfin

The aim is to measure the efficiency and productivity of the Mexican Central Bank (BANXICO) and propose strategies to become more competitive and efficient in the future. Four DEA methods are applied to measure its productivity and efficiency against the European Central Bank, the Reserve Bank of Australia, the Bank of Canada, the Bank of Japan, the Reserve Bank of New Zealand, the Swedish Riksbank, the Swiss National Bank, the Bank of England and the Federal Reserve System of the USA; second, to compare the efficiency of the Latin American Central Banks (LACBs) against the most efficient ones; and third, to assess the efficiency among the LACBs. The input variables are the TIIE (Interbank interest rate of equilibrium) and knowledge of the economy; and the output variables are inflation and the volatility of the GDP (Gross Domestic Product). We obtain that LACB and BANXICO are less efficient than the rest of the banks. We use the weights obtained by the DEA methods to propose economic strategies that BANXICO can implement to be a more efficient central bank compared with the better behavior banks in the sample.

C1609: **The effectiveness of unconventional monetary policiy on risk aversion and uncertainty**
Presenter: **Leonidas Rompolis**, Athens University of Economics and Business, Greece

The impact of ECB's unconventional monetary policy is examined by measuring it by its balance sheet expansion on euro area equity market uncertainty and investors risk aversion within a structural VAR framework. An expansionary balance sheet shock decreases both risk aversion and uncertainty at least in the medium-run. A negative shock on policy rates has also a negative impact on risk aversion and uncertainty. These results are generally robust to different specifications of the VAR model, estimation procedures and identification schemes. Conversely, periods of high uncertainty and risk aversion are followed by a looser conventional monetary policy. The effect of uncertainty and risk aversion on ECBs total assets is not always statistically significant.

CG339 Room Gordon CONTRIBUTIONS IN EMPIRICAL MACROECONOMICS Chair: Jonas Dovern

C0215: **Evaluating the macroeconomic effects of the ECBs unconventional monetary policies**
Presenter: **Sarah Mouabbi**, Banque de France, France
Co-authors: Jean-Guillaume Sahuc

The macroeconomic effects of the European Central Banks unconventional monetary policies is quantified using a dynamic stochastic general equilibrium model which includes a shadow Eonia rate. Extracted from the yield curve, this shadow rate provides an unconstrained measure of the overall stance of monetary policy. Counterfactual analyses show that, without unconventional measures, the euro area would have suffered (i) a cumulative loss of output of around 19% of its pre-crisis level since the Great Recession, (ii) deflation episodes in 2009Q1 and 2016Q1 and (iii) a slowdown in price increases in 2015 and 2016. This translates into year-on-year inflation and GDP growth growth that would have been on average about 0.3% and 0.5% below their actual levels over the period 2014Q1-2016Q1, respectively.

C1546: **The effects of economic policy uncertainty on European economies: Evidence from a TVP-FAVAR**
Presenter: **Alexander Schloesser**, University Duisburg-Essen, Germany
Co-authors: Jan Prueser

Recent events such as the financial and sovereign debt crisis have triggered an increase in European Economic Policy Uncertainty (EPU). We use a TVP-FAVAR model with hierarchical priors on the hyperparameters in order to investigate the effect of EPU on a wide range of macroeconomic variables for eleven European Monetary Union (EMU) countries. First, we find that EPU shocks are transmitted through various channels, such as the real options-,the precautionary savings- and the financial channel. Second, we are able to distinguish between a group of fragile countries (GIIPS-countries) and a group of stable countries (northern countries), where the former are more strongly affected by EPU shocks. Third, while the IRFs for most variables differ only in magnitude and not in sign between groups of countries, responses of long term interest rates to EPU shocks have a different sign across countries. Fourth, we discover that investors and traders react more sensitively than consumers to uncertainty. Fifth, we find that EPU shocks affect monetary policy decisions. Sixth, we provide evidence that the transmission of EPU shocks is quite stable over

time. Finally, the increase in EPU can partly be explained by the state of the European economy and should therefore be treated as an endogenous variable.

C1650: **Empirical analysis on the effect of the fiscal policy in the zero lower bound of nominal short-term interest rate**
Presenter: **Hiroshi Morita**, Hosei University, Japan

For the Japanese economy, the aim is to empirically examine whether the effectiveness of a fiscal policy depends on the stance of monetary policy by comparing the two different approaches which build on the standard time-varying parameters VAR model. In the first approach, the coefficients in the interest rate equation except for its own lags are set to be zero during the period of zero-lower bound (ZLB) to represent the actual situation of ZLB period in which short-term interest rate hardly reacts to any economic variations. We call this model TVP-VAR model with ZLB because it represents the actual situation of ZLB period in which short-term interest rate hardly reacts to any economic variations. In the second approach, we construct TVP-VAR model with a censored variable to compute the implied interest rate when the interest rate hits the ZLB. We call the second model the TVP-VAR model without ZLB, so that this model embodies a counterfactual situation without the ZLB of a nominal interest rate. As a result of comparing fiscal multipliers derived from these two models, we cannot detect the significant evidence that the zero-interest rate policy enhances the effectiveness of the fiscal policy in Japan.

C0478: **Low frequency drivers of the real interest rate: A band spectrum regression approach**
Presenter: **Fabio Busetti**, Banca d'Italia, Italy

An empirical analysis on the underlying drivers of the real interest rate in advanced economies over the last 35 years is presented. We adopt a band spectrum regression approach, which allows to study the link between the real interest rate and its determinants only over low frequencies, leaving aside the business cycle fluctuations and the high frequency noise. Our findings indicate that most of the long-term movements of real interest rates are explained by the evolution of total factor productivity (with a specific role for human capital accumulation) and demography. Monetary policy developments and changes in income inequality on the other hand play a limited role. According to our estimates, over recent years the natural rate of interest remained positive in US and UK but fell below zero in the euro area and Japan.

Monday 18.12.2017 10:50 - 12:55 Parallel Session M – CFE-CMStatistics

EI009 Room Beveridge Hall RECENT ADVANCES IN THE DESIGN OF EXPERIMENTS Chair: Steven Gilmour

E0173: **Bayesian optimal design of experiments: Review, challenges and examples**
Presenter: **David Woods**, University of Southampton, United Kingdom
The design of any experiment is implicitly Bayesian, with prior knowledge being used informally to aid decisions such as which factors to vary and the choice of plausible causal relationships between the factors and measured responses. Adoption of formal Bayesian methods allow uncertainty in these decisions to be incorporated into design selection through prior distributions that encapsulate information available from scientific knowledge or previous experimentation. Further, a design may be explicitly tailored to the aim of the experiment through a decision-theoretic approach with an appropriate loss function. However, finding decision-theoretic optimal designs is challenging, largely due to the typically high-dimensional and intractable integration required to evaluate the expected loss. We review some of the recent research in this area, expand on the some of the challenges in Bayesian design, and present some example solutions.

E0174: **Some advances in designing multi-stratum experiments**
Presenter: **Luzia Trinca**, Unesp, Brazil
Co-authors: Steven Gilmour
Multi-stratum experiments are frequently performed in many areas as, for example, laboratory biology, agriculture and engineering. Often experimental cost does not allow the use of orthogonal layouts and it is important to have general methods to construct efficent designs for the practical problems. The stratum-by-stratum approach is quite flexible and can be used with any design criteria including compound criteria in order to construct flexible and efficient designs. Here we explore the flexibility of such approach, including properties that allow pure error degrees of freedom in the several strata, a common defficiency presented by designs constructed by usual approaches. A few illustrations are given.

E0192: **Geometric orthogonal arrays: Multidimensional space filling property and construction via factor collapse**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Cheng-Yu Sun, Shaowei Cheng
A new class of space-filling designs optimized under a new multi-dimensional space-filling property is introduced which is called *geometric strength*. We propose a systematic construction method via techniques in Galois field for this new class of designs. The factor levels in a regular design are collapsed and the strength of the collapsed design is enhanced. The reversed process to relabel factor levels of the regular design improves its space-filling property. This method is more efficient than the existing methods via level permutations, especially when the number of factor levels is large. When two collapsers are indistinguishable in terms of the strength of the collapsed designs, we propose a new criterion called maximal strength efficiency. It not only maximizes the strength of the collapsed design, but also maximizes the proportion of the projected sub-designs that are full factorials.

EO021 Room CLO B01 MODERN APPLICATIONS OF FUNCTIONAL DATA ANALYSIS Chair: Wenceslao Gonzalez-Manteiga

E1178: **Time-varying proportional odds model for mega-analysis of clustered event times with functional covariates**
Presenter: **Tanya Garcia**, Texas A&M University, United States
Co-authors: Karen Marder, Yuanjia Wang
Mega-analysis, or the meta-analysis of individual data, enables pooling and comparing multiple studies to enhance estimation and power. A challenge in mega-analysis is estimating the distribution for clustered, potentially censored event times where the dependency structure can introduce bias if ignored and functional covariates introduce a secondary complexity. We propose a new proportional odds model with unknown, time-varying coefficients and functional covariates. The model directly captures event dependencies, handles censoring using pseudo-values, and corrects for the impact of functional genetic information. We demonstrate that a simple transformation of the model leads to an easily estimable additive logistic mixed effect model. Our method consistently estimates the distribution for clustered event times even under covariate-dependent censoring and complex functional covariate structures. Applied to three observational studies of Huntington's disease, our method provides, for the first time in the literature, evidence that varying lengths of CAG repeats in the huntingtin gene (modeled as a functional covariate) has very different impacts on motor and cognitive impairments.

E0912: **Constructing probabilistic templates for astronomical lightcurves**
Presenter: **David Jones**, Duke University, United States
Co-authors: Sujit Ghosh, Ana-Maria Staicu, Ashish Mahabal
The lightcurve of an astronomical source (e.g. a star) is its light intensity as a function of time. New telescopes such as the Large Synoptic Survey Telescope (LSST) will observe billions of lightcurves at tens to hundreds of time points, and astronomers want to classify each and make decisions about follow-up observations. Many of the most interesting and informative astronomical sources have periodic lightcurves with shapes that are characteristic of the particular class of source. Given a sufficient number of pre-classified lightcurves we can infer the lightcurve shapes characteristic of different classes (or, more precisely, the distribution of characteristic lightcurve shapes). In this manner we can construct probabilistic templates for lightcurves from each class. These templates can then be input into methods for the key tasks of classifying new sources and optimally scheduling further observations (of sources with few observations). The success of the classification and scheduling methods depends highly on the development of accurate probabilistic templates and we discuss our data driven approach to construct bases that can efficiently capture important lightcurve features. We apply our method to data from the Catalina Real-time Transient Survey.

E1074: **Operator-regularized covariance function estimation for functional data**
Presenter: **Xiaoke Zhang**, George Washington University, United States
Co-authors: Raymond Wong
In functional data analysis (FDA), covariance function is fundamental, not only as a critical quantity for understanding elementary aspects of functional data, but also as an indispensable ingredient for many advanced FDA methods. A new class of nonparametric covariance function estimators is developed in terms of various spectral regularizations of an operator associated with a reproducing kernel Hilbert space. Despite their nonparametric nature, the covariance estimators are automatically positive semi-definite without any additional modification steps. An unconventional representer theorem is established to provide a finite dimensional representation for this class of covariance estimators, which leads to a closed-form expression of the corresponding L^2 eigen-decomposition. Trace-norm regularization is particularly studied to further achieve a low-rank representation, another desirable property which leads to dimension reduction and is often needed in advanced FDA approaches. An efficient algorithm is developed based on the accelerated proximal gradient method. This resulted estimator is shown to enjoy an excellent rate of convergence under both fixed and random designs. The outstanding practical performance of the trace-norm-regularized covariance estimator is demonstrated by a simulation study and the analysis of a traffic dataset.

EO686 Room MAL B20 STOCHASTIC PROCESSES AND EXTREMES Chair: Simone Padoan

E0260: **Predicting extreme influenza epidemics**
Presenter: **Maud Thomas**, Universite Pierre et Marie Curie, France
Co-authors: Holger Rootzen

Influenza viruses are responsible for annual epidemics, causing more than 500,000 deaths per year worldwide. A crucial question for resource planning in public health is to predict the morbidity burden of extreme epidemics. We say that an epidemic is extreme whenever the influenza incidence rate exceeds a high threshold for at least one week. Our objective is to predict whether an extreme epidemic will occur in the near future, say the next couple of weeks. The weekly numbers of influenza-like illness (ILI) incidence rates in France are available from the Sentinel network for the period 1991-2017. ILI incidence rates exhibit two different regimes, an epidemic regime during winter and a non-epidemic regime during the rest of the year. To identify epidemic periods, we use a two-state autoregressive hidden Markov model. A main goal of Extreme Value Theory is to assess, from a series of observations, the probability of events that are more extreme than those previously recorded. Given the autoregressive structure of the data, we choose to fit one of the multivariate generalized Pareto distribution models. For these models, explicit densities are given, and formulas for conditional probabilities can then be deduced, from which we can predict if an epidemic will be extreme, given the first weeks of observation.

E1184: **Some results on joint record events**
Presenter: **Amir Khorrami Chokami**, Universita Bocconi, Italy
Co-authors: Michael Falk, Simone Padoan

Let $X_1, X_2, \ldots$ be i.i.d. copies of a r.v. $X \in \mathbb{R}^d$ with a continuous joint distribution function F. When $d = 1$, the stochastic behavior of the sequence of subsequent records is well known. Differently, we study the stochastic behavior of arbitrary $X_j, X_k, j < k$, under the condition that they are records, without knowing their order in the sequence of records. The results are completely different. In particular, it turns out that the distribution of X_k, being a record, is not affected by the additional knowledge that X_j is a record as well. The distribution of X_j, being a record, is however affected by the additional knowledge that X_k is a record as well. If F has a density, then the gain of this additional information, measured by the corresponding Kullback-Leibler distance, is independent of F.In the multivariate case, a random vector is defined a complete record if each component is a univariate record. In the case of independent components, it is known that there are finitely many complete records. New results on the arrival times of complete records are provided, and it is investigated the distribution of the terminal record.

E0768: **The extremal index for IGARCH(p,q) processes with skewed innovations**
Presenter: **Fabrizio Laurini**, University of Parma, Italy

GARCH process are widely used for modelling features commonly found in observed financial returns. The extremal properties of these processes are of wide interest for market risk management. Only for simple GARCH(1,1) extremes have been fully characterised, and much remains to be found about the dependence structure. In particular, the mean number of extreme values in a short term cluster, i.e., the reciprocal of the extremal index, has only been characterised in special cases which exclude all GARCH(p,q) processes that are used in practice, e.g., with innovations with unbounded support or asymmetry. Although recent research has identified the multivariate regular variation property of stationary GARCH(p,q) processes, currently there are no methods for numerically evaluating key components of these characterisations. We overcome these issues and are able to generate the forward tail chain of the process to derive the extremal index even for the general Integrated GARCH(p,q), considering unbounded and asymmetric innovations. The convergence of our numerical algorithm is very fast due to a efficient implementation of a particle filtering simulation technique.

E0652: **Exceedance-based nonlinear regression of residual dependence in extremes**
Presenter: **Linda Mhalla**, University of Geneva, Switzerland
Co-authors: Valerie Chavez-Demoulin, Thomas Opitz

Most environmental processes are spatial by nature. For such spatial processes, different tail behavior types ranging from asymptotic independence to asymptotic dependence are observed. One way of characterizing different tail decay behavior as events become more extreme is via the Pickands dependence function and the angular dependence function, which are appropriate in the presence of extremal dependence and independence, respectively. Motivated by the analysis of the dependence between high background concentrations of nitrogen dioxide measured at different sites in France, we develop a semiparametric framework to estimate both the Pickands and the angular dependence functions based on the excesses and deficits of a min-projection univariate random variable, respectively. Moreover, the effect of a set of predictors on the dependence functions is taken into account based on a generalized additive modeling framework. The methodology is applied to capture the behavior through time and as a function of the distance between sites of the tail dependence between nitrogen dioxide measurements.

E1257: **Human life is unlimited – but short**
Presenter: **Holger Rootzen**, Chalmers, Sweden
Co-authors: Dmitrii Zholud

Does the human lifespan have an impenetrable biological upper limit which ultimately will stop further increase in life lengths? Answers to this question are important for our understanding of the ageing process, and for the organization of society, and have led to intense controversies. Demographic data for humans have been interpreted as showing existence of a limit close to the age, 122.45 years, of the longest living documented human, Jeanne Calment, or even as indication of a decreasing limit, but also as evidence that a limit does not exist. The aim is to study what data says about human mortality after age 110. We show that in north America, Australia, western Europe, and Japan the yearly probability of dying after age 110 is constant and about 53% per year. Hence there is no finite limit to the human lifespan. Still, given the present stage of biotechnology, it is unlikely that during the next 25 years anyone will live longer than 128 years in these countries. Data, remarkably, show little difference in mortality after age 110 between men and women, between earlier and later periods, between ages, or between persons with different lifestyles or genetic backgrounds. These results can help testing biological theories of ageing and aid early confirmation of success of efforts to find a cure for ageing.

EO431 Room MAL B30 ADVANCES IN STATISTICAL ANALYSIS OF COMPLEX FUNCTIONAL DATA Chair: Jian Qing Shi

E1104: **Analysis of spatially correlated functional data objects**
Presenter: **Salihah Alghamdi**, University of Glasgow, United Kingdom
Co-authors: Surajit Ray

Datasets observed over space and time are becoming increasingly abundant and researchers use methods ranging from spatially correlated time series to functional data analysis. There exists a host of methods for modelling using spatially correlated functional data analysis techniques, most of which were motivated by the specific application areas. However, there does not exist any clear pathway to determining the best methodology for a new application. This motives us to present a flow chart of the expected process that one should follow when analysing space-time data. We will demonstrate the use of this framework using a case study, which models brain data.

E0707: **A time-dependent PDE regularization to model functional data defined over spatio-temporal domains**
Presenter: **Eleonora Arnone**, Politecnico di Milano, Italy
Co-authors: Laura Azzimonti, Fabio Nobile, Laura Sangalli
A new method is proposed for the analysis of functional data defined over spatio-temporal domains. These data can be interpreted as time evolving surfaces or spatially dependent curves. The proposed method is based on regression with differential regularization. We are in particular interested in the case when prior knowledge on the phenomenon under study is available. The prior knowledge is described in terms of a time-dependent Partial Differential Equation (PDE) that jointly models the spatial and temporal variation of the phenomenon. We consider various samplings designs, including geo-statistical and areal data. We show that the corresponding estimation problem are well posed and can be discretized in space by means of the Finite Element method, and in time by means of the Finite Difference method. The model can handle data distributed over spatial domains having complex shapes, such as domains with strong concavities and holes. Moreover, various types of boundary conditions can be considered. The proposed method is compared to existing techniques for the analysis of spatio-temporal models, including space-time kriging and methods based on thin plate splines and soap film smoothing. As a motivating example, we study the blood flow velocity field in the common carotid artery, using data from Echo-Color Doppler.

E0839: **Modelling covariance structure of multi-dimensional function-valued processes using convolved Gaussian processes**
Presenter: **Evandro Konzen**, Newcastle University, United Kingdom
Co-authors: Jianqing Shi
In order to consider the cross-covariance structure among several multi-dimensional functions, we use convolved Gaussian processes ensuring that the covariance matrix for the multivariate functional response is positive definite. We discuss implementation issues and show examples based on simulated data. An application to human fertility data is also illustrated.

E0838: **Functional data in forensic entomology: Estimation of temperature dependent growth processes**
Presenter: **Davide Pigoli**, King's College London, United Kingdom
Larvae (or maggots) collected at crime scenes contribute important pieces of information to police investigations. In particular, their hatching time provides a lower bound for the post mortem interval, i.e. the interval between death and the discovery of the body. A functional data analysis approach is described here to model the local growth rate from the experimental data on larval development, where larvae have been exposed to a relatively small number of constant temperature profiles. This allows us to reconstruct varying temperature growth profiles and use them to find out the most likely hatching time for a sample of larvae from the crime scene.

E1665: **Effects of changing the reference measure in statistical processing of density functions**
Presenter: **Renata Talska**, Palacky University Olomouc, Czech Republic
Co-authors: Alessandra Menafoglio, Karel Hron
Data resulting from surveys frequently occurs in form of discrete distributional observations which can be subsequently represented by probability density functions (PDFs). This motivates an increasing interest in statistical tools for their statistical processing. Although functional data analysis (FDA) consists of a wide range of such tools, they are typically designed in L^2 space, thus cannot be directly applied to densities, as the metrics of L^2 does not honor their geometric properties. This has recently motivated the construction of the so-called Bayes Hilbert spaces, which result from the generalization to the infinite dimensional setting of the Aitchison geometry for compositional data. More precisely, when focusing on PDFs with bounded domain $I \subset R$, one can consider the Bayes space of positive real functions with logarithm which is square-integrable with respect to Lebesgue reference. Nonetheless, for unbounded supports, different reference measures need to be used. The aim is to show the effects of changing the reference Lebegue measure to a general probability measure, with the emphasis on its practical implications for the Simplicial Functional Principal Component Analysis (SFPCA) which has been recently designed for dimension reduction of PDFs using the Bayes space methodology.

EO146 Room MAL B35 RECENT DEVELOPMENT IN NEUROIMAGING RESEARCH Chair: Tingting Zhang

E0177: **Recent developments in Bayesian modelling of brain dynamics**
Presenter: **Peter Zeidman**, University College London, United Kingdom
Co-authors: Karl Friston
Some recent developments in modelling brain activity and connectivity for cognitive neuroscientists are introduced. These methods enable experimenters to specify forward models which describe how brain circuitry gives rise to imaging data e.g. fMRI, EEG, MEG. To test hypotheses, these models can be fitted to the data and compared based on their evidence. We will provide an overview of novel methods for group connectivity studies: Parametric Empirical Bayes, PEB, and Bayesian Model Reduction, BMR. These tools can, for example, be used to distinguish patients from controls based on their brain connectivity, or to predict clinical scores. We will also describe a recent application of these methods for Bayesian fusion a multi-modal neural circuitry model which may clarify the relationship between fMRI, EEG and MEG data. Together, these methods may offer new ways to test interesting hypotheses about the brain, drawing on data from multiple imaging modalities.

E0207: **Bayesian inference of the brain's effective connectivity**
Presenter: **Tingting Zhang**, University of Virginia, United States
A new high-dimensional dynamic model is built for the directional interaction, also called effective connectivity, among brain regions. The proposed model is based on a multivariate autoregressive model for time series measurements of brain activity. In order to distinguish strong connections from weak ones, we impose sparsity on the model parameters for the effective connectivity among every pair of brain regions. Specifically, the new model features a cluster structure, which consists of modules of densely connected brain regions. We show that the autoregressive model can outperform a high-dimensional ordinary differential equation model for the brain's effective connectivity. We develop a unified Bayesian framework to make inferences about the brain networks using a stochastic block prior for network structures. We apply the proposed method to time series data of brain activity, specifically, intracranial EEG data, and investigate brain network changes around time.

E0613: **Geometry and registration in functional neuroimaging studies**
Presenter: **John Aston**, University of Cambridge, United Kingdom
Co-authors: Eardi Lila
When imaging data is collected from different subjects for comparison or analysis across the population, there is almost always some preprocessing done before analysis. One of the most common steps is some form of registration (also known as spatial normalization). Here data is (non-)linearly warped geometrically from one domain to another, often taking with it some additional, functional information. However, it is routinely the case, that once the registration takes place, this warping information is discarded. We will discuss approaches where this information, along with the registered functional data both form integral parts of the modelling. We will investigate how shape might be related to function via these models, in situations as diverse as musculoskeletal imaging through to neuroimaging. It will be shown that by including both components in the model, a much fuller description of the data can be achieved.

E0618: **White matter fiber estimation, smoothing and tracking**
Presenter: **Jie Peng**, University of California Davis, United States
The estimation of local neuronal fiber organizations is investigated based on diffusion MRI data and apply fiber tracking algorithms to reconstruct white matter fiber tracts. We apply our methods to the Alzheimer's Disease Neuroimaging Initiative and the Human Connectome Project data. We also discuss various issues and impact of experimental parameters in fiber estimation, smoothing and tracking.

E1256: **Calcium imaging: State-of-the-art and future challenges**
Presenter: **Jordan Rodu**, University of Virginia, United States
Calcium imaging is a powerful technique for observing neuronal populations of ever-increasing sizes, and it creates a unique opportunity for studying population-level dynamics in the brain. Despite the rapid advancement of the technique, many statistical challenges remain. For instance, there is substantial noise introduced in the recordings; images are obtained at a relatively low sampling rate; and there is a nonlinear relationship between the calcium fluorescence and action potentials. These and other factors make spike-timing inference difficult and susceptible to artifacts from pre-processing. We identify major areas of focus and briefly touch on the current state-of-the-art solutions. We also discuss new statistical challenges for calcium imaging studies.

EO140 Room Bloomsbury STATISTICAL INFERENCE USING ELECTRONIC HEALTH RECORDS Chair: Michael Daniels

E0230: **Statistical modeling for heterogeneous populations with application to hospital admission prediction**
Presenter: **Menggang Yu**, University of Wisconsin - Madison, United States
The motivation arises from risk modeling for large hospital and health care systems that provide services to diverse and complex patients. Modeling such heterogeneous populations poses many challenges. Often, heterogeneity across a population is determined by a set of factors such as chronic conditions. When these stratifying factors result in overlapping subpopulations, it is likely that the covariate effects for the overlapping groups have some similarity. We propose to exploit this similarity by imposing structural constraints on the importance of variables in predicting outcome such as hospital admission. We prove an oracle property for our estimation method that enables construction of confidence intervals. We also show that even when the structural assumptions are misspecified, our method will still include all of the truly nonzero variables in large samples and therefore provide valid asymptotic statistical inference. We demonstrate impressive performance of our method in extensive numerical studies and on an application in hospital admission prediction and validation for the Medicare population at the University of Wisconsin-Madison Health System.

E0421: **Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data**
Presenter: **Jenny Haggstrom**, Umea University, Sweden
Co-authors: Cheng Ju, Richard Wyss, Jessica M Franklin, Sebastian Schneeweiss, Mark van der Laan
Propensity score (PS) based estimators are increasingly used for causal inference in observational studies. However, model selection for PS estimation in high-dimensional data has received little attention. In these settings, PS models have traditionally been selected based on the goodness-of-fit for the treatment mechanism itself, without consideration of the causal parameter of interest. Collaborative minimum loss-based estimation (C-TMLE) is a novel methodology for causal inference that takes into account information on the causal parameter of interest when selecting a PS model. This "collaborative learning" considers variable associations with both treatment and outcome when selecting a PS model in order to minimize a bias-variance trade off in the estimated treatment effect. We introduce a novel approach for collaborative model selection when using the LASSO estimator for PS estimation in high-dimensional covariate settings. To demonstrate the importance of selecting the PS model collaboratively, we designed quasi-experiments based on a real electronic healthcare database, where only the potential outcomes were manually generated, and the treatment and baseline covariates remained unchanged. Results showed that the C-TMLE algorithm outperformed other competing estimators for both point estimation and confidence interval coverage.

E0572: **Adjusting for selection bias due to missing data in EHR-based research**
Presenter: **Sebastien Haneuse**, Harvard TH Chan School of Public Health, United States
Co-authors: Sarah Peskoe, David Arterburn, Michael Daniels
While electronic health records (EHR) data provide unique opportunities for medical research, there are numerous challenges that must be dealt with. Among these, selection bias due to missing data is under-appreciated. While standard missing data methods are often applied in the EHR context, they will, in general, fail to capture the complexity of the data so that residual selection bias may remain. Building on a recently-proposed framework for characterizing how data arise in EHR-based studies, we develop and evaluate a statistical framework for regression modeling based on inverse probability weighting that adjusts for selection bias in the complex setting of EHR-based research. We show that the resulting estimator is consistent and asymptotically Normal, and derive the form of the asymptotic variance. Plug-in estimators for the latter are proposed. We use simulations to: (i) highlight the potential for bias in EHR studies when standard approaches are used to account for selection bias, and (ii) evaluate the small-sample operating characteristics of the proposed framework. Finally, the methods are illustrated using data from an on-going EHR-based study of bariatric surgery on BMI.

E0745: **Bayesian functional clustering for laboratory data from electronic medical records**
Presenter: **Jason Roy**, University of Pennsylvania, United States
Co-authors: Bret Zeldow
A Bayesian semiparametric mixed model is proposed for longitudinal data by using an enriched Dirichlet process (EDP) prior. To account for nonlinearities in the outcome over time, we use splines to model the time effect. The nonparametric EDP prior is placed on the regression and spline coefficients, the error variance, and the parameters governing the predictor space. The goal is to predict the outcome at unobserved time points for subjects with outcome data at other time points and for completely new subjects. We find improved prediction over mixed models with Dirichlet process (DP) priors when there are a large number of predictors. Our method is demonstrated with electronic health records consisting of new initiators of second generation antipsychotics, which are known to increase the risk of diabetes. We use our model to predict laboratory values indicative of diabetes for each individual and assess incidence of suspected diabetes from the predicted dataset.

E1738: **Classification of heart conditions using functional data analysis**
Presenter: **Chibueze Ogbonnaya**, University of Nottingham, United Kingdom
Co-authors: Simon Preston, Andrew Wood, Karthik Bharath
A functional data analysis approach to heart defect detection using heart signals recorded by electrocardiograms (ECGs) is proposed. ECGs can be thought of as continuous functions having an amplitude and phase component. Raw heart signals are usually noisy with artifacts such as baseline wander. When comparing two or more ECGs, there are also issues with arbitrary location and scale. To remove these artifacts and deal with the issues encountered when comparing two or more signals, we propose amplitude registration models and give closed form solutions for the amplitude parameters. For some heart conditions which are characterised by amplitude changes, such as high peaks or inverted curves, to classify the subjects, we first perform functional principal component analysis (FPCA) on the registered functions and use the PC scores as predictors in a classifier. When heart conditions correspond to phase changes, we propose to use a parametric family of warping functions to detect these phase differences. Classification is done using the estimated parameters as predictors in a classifier. The predictive accuracy of our method using

leave-one-out cross-validation are 93% and 100% for the amplitude classification of Myocardial Infarction and Cardiomyopathy respectively. This compares favourably with existing approaches for classification of ECGs.

EO386 Room Chancellor's Hall VAST TIME SERIES ANALYSIS AND APPLICATIONS Chair: Clifford Lam

E0470: **Simultaneous multiple change-point and factor analysis for high-dimensional time series**
Presenter: **Piotr Fryzlewicz**, London School of Economics, United Kingdom
Co-authors: Matteo Barigozzi, Haeran Cho

The first comprehensive treatment of high-dimensional time series factor models with multiple change-points in their second-order structure is proposed. We operate under the most flexible definition of piecewise stationarity, and estimate the number and locations of change-points consistently as well as identifying whether they originate in the common or idiosyncratic components. Through the use of wavelets, we transform the problem of change-point detection in the second-order structure of a high-dimensional time series, into the (relatively easier) problem of change-point detection in the means of high-dimensional panel data. Our methodology circumvents the difficult issue of the accurate estimation of the true number of factors by adopting a screening procedure. In extensive simulation studies, we show that factor analysis prior to change-point detection improves the detectability of change-points, and identify and describe an interesting 'spillover' effect in which substantial breaks in the idiosyncratic components get, naturally enough, identified as change-points in the common components, which prompts us to regard the corresponding change-points as also acting as a form of 'factors'. We introduce a simple graphical tool for visualising the piecewise stationary evolution of the factor structure over time. Our methodology is implemented in the R package factorcpt, available from CRAN.

E0995: **The local partial autocorrelation function and its application to forecasting**
Presenter: **Marina Knight**, University of York, United Kingdom
Co-authors: Rebecca Killick, Guy Nason, Idris Eckley

The regular and partial autocorrelation functions are powerful tools for stationary time series modelling and analysis. However, in many applied situations time series are not stationary and in these settings the use of the regular (classical) and partial autocorrelations can give misleading answers. We introduce the local partial autocorrelation function and establish the asymptotic behaviour of its estimators. We demonstrate its practical utility as a tool in economic applications and in addition, we propose its use to improve the forecasting of locally stationary time series. Our new forecasting method exhibits excellent forecasting and prediction interval coverage for simulated nonstationary data, as well as for economic and wind energy data.

E0781: **Regularised forecasting via smooth-rough partitioning of the regression coefficients**
Presenter: **HyeYoung Maeng**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz

A way of modelling temporal dependence in random functions $X(t)$ in the framework of linear regression is introduced. Based on discretised curves $(X_i(t_0), X_i(t_1), ..., X_i(t_T))$, the final point $X_i(t_T)$ is predicted from $(X_i(t_0), X_i(t_1), ..., X_i(t_{T-1}))$. The proposed model flexibly reflects the relative importance of predictors by partitioning the regression parameters into a smooth and a rough regime. Specifically, unconstrained (rough) regression parameters are used for influential observations located close to $X_i(t_T)$, while the set of regression coefficients for the predictors positioned far from $X_i(t_T)$ are assumed to be sampled from a smooth function. This both regularises the prediction problem and reflects the 'fading memory' structure of the time series. The point at which the change in smoothness occurs is estimated from the data via a technique akin to change-point detection. The joint estimation procedure for the smoothness change-point and the regression parameters is presented, and the asymptotic behaviour of the estimated change-point is analysed. The usefulness of the new model is demonstrated through simulations and two real data examples, involving stock volatility series and country mortality data.

E0589: **Spatial lag model with time-lagged effects and spatial weight matrix estimation**
Presenter: **Cheng Qian**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

A spatial lag model is considered which has different spatial weight matrices for different time-lagged spatial effects, while allows both the sample size T and the panel dimension N to grow to infinity together. To overcome potential misspecifications of these spatial weight matrices, we estimate each one by a linear combination of a set of M specified spatial weight matrices, with M being finite. Moveover, by penalizing on the coefficients of these linear combinations, oracle properties for these penalized coefficient estimators are proved, including their asymptotic normality and sign consistency. Other parameters of the model are estimated by profile-least square type of estimators after introducing covariates which serve similar functions as instrumental variables. Asymptotic normality for our estimators are developed under a framework of functional dependence, which is a measure of time series dependence. The proposed methods are illustrated using both simulated and real financial data.

EO640 Room Court ADVANCES IN THEORY AND MODELING OF SPATIAL AND SPATIO-TEMPORAL DATA Chair: Chae Young Lim

E0351: **Score function approximation for Gaussian processes using equivalent kernels**
Presenter: **Zachary Mullen**, University of Colorado, United States
Co-authors: William Kleiber

Large, spatially indexed datasets bring with them a host of computational and mathematical challenges. Parameter estimation often relies on maximum likelihood, which, for Gaussian processes, involves matrix manipulations of the covariance matrix including solving systems of equations and determinant calculations. The score function, on the other hand, avoids direct calculation of the determinant, but still requires solving a large number of linear equations. We propose an equivalent kernel approximation to the score function of a Gaussian process. A nugget effect is required for the approximation, and we find the nugget must not be negligible for adequate approximations. For large, noisy processes, the approximation is fast, accurate, and compares well against existing approaches.

E0552: **Isotropic variogram models on all spheres**
Presenter: **Juan Du**, Kansas State University, United States

Variogram or variogram matrix functions play an important role in modeling dependence structure among multiple processes at different locations in spatial statistics. The aim is to characterize the variogram models on spheres of all dimensions for intrinsically stationary, isotropic, univariate, and multivariate processes. Efficient approaches are proposed to construct a variety of isotropic variogram functions including simple polynomial structures. Series representations and spherical behavior of intrinsic stationary random fields are elaborated in both theoretical and simulation study. Applications of the proposed model and related theoretical results are demonstrated using simulation and real data analysis.

E0917: **Variable selection with spatially autoregressive errors: A generalized moments LASSO Estimator**
Presenter: **Arnab Bhattacharjee**, Heriot-Watt University, United Kingdom

A generalized moments LASSO estimator, combining LASSO with GMM, is proposed for penalized variable selection and estimation under the spatial error model with spatially autoregressive errors. We establish parameter consistency and selection sign consistency of the proposed estimator

in the low dimensional setting when the parameter dimension $p <$ sample size n, as well as the high dimensional setting with p greater than and growing with n. Finite sample performance of the method is examined by simulation, compared against the LASSO for IID data. The methods are applied to estimation of a spatial Durbin model for the Aveiro housing market (Portugal).

E0808: **Tail estimation for the cross-spectral density of a bivariate stationary Gaussian random field**
Presenter: **Wei-Ying Wu**, National Dong Hwa University, Taiwan
Consider the bivariate stationary Gaussian random field model. Under some assumptions on high-frequency behavior of (cross-)spectral densities, a modified multivariate Whittle likelihood are proposed to estimate parameters, which dominate tail behaviors of the (cross-)spectral densities. Under the regular infill sampling, we discuss the asymptotic properties of the proposed estimators as well as simulation results.

EO057 Room G11 STATISTICAL ANALYSIS OF COMPLEX DATA Chair: Antonio Cuevas

E0418: **Nonparametric estimation of a latent variable model - a new approach**
Presenter: **Tim Fabian Schaffland**, University of Tuebingen, Germany
Co-authors: Augustin Kelava, Michael Kohler, Adam Krzyzak
A new nonparametric latent variable approach is presented. The model is estimated without specifying the underlying distributions of the latent variables. In a first step, we fit a common factor analysis model to the observed variables. The main trick in estimation of the common factor analysis model is to estimate the values of the latent variables in such a way that the corresponding empirical distribution asymptotically satisfies the conditions that characterize the distribution of the latent variables uniquely. In a second step, we apply suitable nonparametric regression techniques to analyze the relation between the latent variables in this model. Theoretical results (e.g., concerning consistency of the estimates) are briefly presented. Furthermore, the finite sample size performance of the proposed approach is illustrated by applying it to simulated data in simulation studies.

E0542: **Bayesian two-level model for partially ordered repeated responses**
Presenter: **Xiaoqing Wang**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Xiangnan Feng, Xinyuan Song
Owing to the questionnaire design and problem nature, partially ordered data that are neither completely ordered nor completely unordered are frequently encountered in behavioral and medical researches. However, among literature, little attention has focused on longitudinal observations with partially ordered data structure. We propose a Bayesian two-level regression model for analyzing longitudinal data with partially ordered responses. The first-level model is defined for partially ordered observations taken at each time point nested within individual and the second-level model is defined for individuals to assess the effect of their characteristics on longitudinal responses. A full Bayesian approach with Markov Chain Monte Carlo (MCMC) algorithm is developed for statistical inference. A simulation study demonstrates that the developed methodology performs satisfactorily. An application to a longitudinal study concerning adolescent substance use is presented.

E0636: **Consistent test on semiparametric functional coefficient models with integrated covariates**
Presenter: **Shan Dai**, The Chinese University of HongKong, Hong Kong
Cointegration is a powerful tool in studying long-run relationships among integrated time series and is widely used in macroeconomics and financial time series analysis. Nonetheless, empirical evidence often fails to support the existence of cointegrating relations with fixed cointegrating slope coefficients. Due to plausible source of variability into the cointegrating vectors, semiparametric functional coefficient cointegration model attracts wide attention recently. Based on the asymptotic result for both the regressors and covariates are I(1) variables, we propose consistent test for the instability of cointegrating parameters in the present framework. The asymptotic distributions of test statistics under both null and alternative hypotheses are studied. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed test. Furthermore, we use the proposed test to help verify the test result for purchasing power parity hypothesis between U.S. and Canadian dollars, which is consistent with some recent result.

E0779: **Bayesian quantile scale on image regression**
Presenter: **Qi Yang**, Department of Statistics, the Chinese University of Hong Kong, Hong Kong
Co-authors: Xinyuan Song
Imaging data are very common in substantive research, especially in medical studies. The analysis of imaging data can reveal the relationship between the images collected and clinical outcomes of interest. We propose a quantile scale on image regression model to provide a comprehensive analysis of the relationship between a scalar response and imaging predictors. The high dimension of imaging data can be reduced with the aid of efficient functional principle component analysis (FPCA) method. A Bayesian approach together with Markov chain Monte Carlo algorithm is developed to conduct statistical inference. Simulation results demonstrate that the proposed model performs satisfactorily in finite samples. A real example is offered to illustrate the usefulness of the proposed model.

E1460: **Robust inference for subgroup analysis with general transformation models**
Presenter: **Wenxin Liu**, The Chinese University of Hong Kong, Hong Kong
A crucial step in developing personalized treatment strategies is to identify the subgroups of patients of a heterogeneous population. We consider a general class of heterogeneous transformation models for subgroup identification, under which an unknown monotonic transformation of the response is linearly related to the covariates via subject-specific regression coefficients with unknown error distribution and unknown *priori* grouping information. This class of models is broad enough to cover many popular models, including some novel heterogeneous linear models. We propose a robust method based on the maximum rank correlation and a concave fusion to automatically identify the subgroup structure and estimate the subgroup-specific treatment effects simultaneously. We establish the theoretical properties of our proposed estimate under regularity conditions. A random weighting resampling scheme is used for variance estimation. The proposed procedure can be easily extended to handle censored data.Numerical studies including simulations and a real data analysis demonstrate that the proposed method performs reasonably well in practical situations.

EO409 Room G3 SCALABLE METHODS FOR COMPLEX DATA Chair: Po-Ling Loh

E1342: **On approximation guarantees for greedy low rank optimization**
Presenter: **Sahand Negahban**, Yale University, United States
New approximation guarantees are provided for greedy low rank matrix estimation under standard assumptions of restricted strong convexity and smoothness. Our novel analysis also uncovers previously unknown connections between the low rank estimation and combinatorial optimization, so much so that our bounds are reminiscent of corresponding approximation bounds in submodular maximization. Additionally, we provide also provide statistical recovery guarantees. Finally, we present empirical comparison of greedy estimation with established baselines on two important real-world problems.

E1381: **High-dimensional classification with positive and unlabeled data**
Presenter: **Garvesh Raskutti**, University of Wisconsin-Madison, United States

In a number of scientific technologies, a classification problem presents only positive and unlabeled data. For example, deep mutational scanning in biochemistry is a high-throughput technology that relates biochemical function to sequence structure and this technology often only provides positive (indicating functional) and unlabeled responses for different protein sequences. Unlabeled data presents a challenge in classification which typically leads to a non-convex optimization problem, since there are hidden variables to indicate whether the unlabeled responses are positive or negative. Furthermore, since the protein sequences are long, the total number of features or covariates in the space is large. We present an approach which combines EM algorithm combined with quadratic majorization to address the computational challenge associated with high-dimensional PU learning. Furthermore, we provide statistical guarantees which prove convergence to a local minima in the high-dimensional setting. The performance of the algorithm is demonstrated both on simulated data and a real-world problem which addresses the question of how protein sequence structure influences biochemical function.

E1772: **A general framework for uncovering dependence networks**
Presenter: **Johannes Lederer**, University of Washington, United States

Dependencies in multivariate observations are a unique gateway to uncovering relationships among processes. An approach that has proved particularly successful in modeling and visualizing such dependence structures is the use of graphical models. However, whereas graphical models have been formulated for finite count data and Gaussian-type data, many other data types prevalent in the sciences have not been accounted for. For example, it is believed that insights into microbial interactions in human habitats, such as the gut or the oral cavity, can be deduced from analyzing the dependencies in microbial abundance data, a data type that is not amenable to standard classes of graphical models. We present a novel framework that unifies existing classes of graphical models and provides other classes that extend the concept of graphical models to a broad variety of discrete and continuous data, both in low- and high-dimensional settings. Moreover, we present a corresponding set of statistical methods and theoretical guarantees that allows for efficient estimation and inference in the framework.

E1776: **Posterior concentration for Bayesian regression trees and their ensembles**
Presenter: **Veronika Rockova**, University of Chicago, United States
Co-authors: Stephanie van der pas

Since their inception in the 1980's, regression trees have been one of the more widely used non-parametric prediction methods. Tree-structured methods yield a histogram reconstruction of the regression surface, where the bins correspond to terminal nodes of recursive partitioning. Trees are powerful, yet susceptible to over-fitting. Strategies against overfitting have traditionally relied on pruning greedily grown trees. The Bayesian framework offers an alternative remedy against overfitting through priors. While the consistency of random histograms, trees and their ensembles has been studied quite extensively, the theoretical understanding of the Bayesian counterparts has been missing. We take a step towards understanding why/when do Bayesian trees and their ensembles not overfit. To address this question, we study the speed at which the posterior concentrates around the true smooth regression function.We propose a spike-and-tree variant of the popular Bayesian CART prior and establish new theoretical results showing that regression trees (and their ensembles) (a) are capable of recovering smooth regression surfaces, achieving optimal rates up to a log factor, (b) can adapt to the unknown level of smoothness and (c) can perform effective dimension reduction when $p > n$. These results provide a piece of missing theoretical evidence explaining why Bayesian trees (and additive variants thereof) have worked so well in practice.

E1510: **Random forests for big data**
Presenter: **Jean-Michel Poggi**, University Paris-Sud Orsay, France
Co-authors: Robin Genuer, Nathalie Villa-Vialaneix, Christine Tuleau-Malot

Big Data is one of the major challenges of statistical science and has numerous consequences from algorithmic and theoretical viewpoints. Big Data always involve massive data, but they also often include online data and data heterogeneity. Recently some statistical methods have been adapted to process Big Data, like linear regression models, clustering methods and bootstrapping schemes. Based on decision trees combined with aggregation and bootstrap ideas, random forests were introduced in 2001. They are a powerful nonparametric statistical method allowing to consider in a single and versatile framework regression problems, as well as two-class and multi-class classification problems. Focusing on classification problems, available proposals that deal with scaling random forests to Big Data problems are selectively reviewed. These proposals rely on parallel environments or on online adaptations of random forests. We also describe how the out-of-bag error is addressed in these methods. Then, we formulate various remarks for random forests in the Big Data context. Finally, we experiment five variants on two massive datasets, a simulated one as well as a real-world dataset. These numerical experiments lead to highlight the relative performance of the different variants, as well as some of their limitations.

EO128 Room G4 ADVANCES IN SURVIVAL AND RELIABILITY Chair: Juan Eloy Ruiz-Castro

E0286: **Statistical testing of availability for mining technological systems**
Presenter: **Jaromir Antoch**, Charles University, Czech Republic

Conducting high-level reliability testing of any complicated technological system under real operational conditions requires a significant amount of resources as time, materials, human skills, etc. One of the possible ways to increase effectiveness is shortened (accelerated) testing. However, this shortening of reliability testing should not compromise the precision of the estimated reliability parameters. We develop a statistical model showing that even environmentally sustainable mining can be profitable. Our basic technique is distribution theory applied to productivity criteria. Two generic working scenarios have been obtained. We illustrate the model on two currently active mining sites, the Chuquicamata copper mine in Chile and the opencast coal mine Libous in the Czech Republic. We show that under very realistic conditions for both countries, the Czech Republic and Chile mining companies can afford to stop mining activities for high thresholds of air pollutants without a substantial loss of productivity.

E0317: **Modeling a complex multi-state warm standby system with loss of units through a D-MMAP**
Presenter: **Mohammed Dawabsha**, University of Granada, Spain
Co-authors: Juan Eloy Ruiz-Castro

In reliability literature it is usual to consider the replacement of a unit which has undergone a non-repairable failure in a negligible unit of time. A complex multi-state warm standby system subject to different types of events, internal and accidental external failures, inspections and preventive maintenance is modeled by considering a discrete Markovian arrival process with marked arrivals. The system is composed of a finite number of units, including the main one and the others in a warm standby. When a non-repairable failure occurs and the system can keep on working, the unit is removed and the system continues working with one unit less. If all units are removed, the system is then reinitialized by considering a new identical one. The number of repairpersons depends on the number of units in the system. Inspections have place randomly and if the observed online unit is worn out, then it goes to preventive maintenance. Different repair times are considered for corrective repair of the online unit, warm standby units and preventive maintenance. The system is modeled and the transient distribution is determined. Some reliability measures of interest are calculated. Rewards and costs are included. A numerical example illustrates the model.

E0493: **Non-parametric estimators for estimating bivariate survival function under randomly censored and truncated data**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Hongsheng Dai, Huan Wang
In bivariate survival analysis it is common to dealt with incomplete information of the data, due to random censoring and random truncation. Such kind of data occurs in many research areas, such as medicine, economics, insurance and social sciences. Most existing research papers focused on bivariate survival analysis when components are either censoring or truncation or where one component is censored and truncated, but the other one is fully observed. Bivariate survival function estimation when both components are censored and truncated has received considerable attention recently. These methods, however, used an iterative computing method which is computationally heavy. Some authors proposed an estimator based on a polar coordinate transformation, which does not require iterative calculations and its large sample properties are established. Starting from their paper, we extend their methods to a class of estimators, based on different data transformations. In particular assuming that the components are both random truncation and random censoring, we propose a class of nonparametric estimators for the bivariate survival function. The proposed class of nonparametric estimators uses such parametric information via a data transformation approach and thus provides more accurate estimates than existing methods without using such information. The proposed method is also justified via a simulation study and an application on an economic data set.

E1405: **Joint models for survival and multivariate longitudinal data**
Presenter: **Mariangela Zenga**, Universita' degli Studi di Milano-Bicocca -DISMEQ, Italy
Co-authors: Marcella Mazzoleni
The joint models analyse the effect of longitudinal covariates onto the risk of one or more events. They are composed of two sub-models, the longitudinal and the survival sub-model. For the longitudinal sub-model a multivariate mixed model can be proposed, considering fixed and random effects. Whereas for the survival sub-model, a Cox proportional hazards model is proposed, considering jointly the influence of two longitudinal covariates onto the risk of the event. The purpose is to extend an estimation method based on a joint likelihood formulation to the case in which the longitudinal sub-model is multivariate. The parameters estimation is based on the maximisation of the likelihood function achieved through the implementation of an Expectation-Maximisation (EM) algorithm. In the M-step a one-step Newton-Raphson update is used, as for some parameters estimators, it is not possible to obtain closed-form expression. In addition, a Gauss-Hermite approximation is applied for some of the integrals involved.

E1462: **A generic framework for recurrent event data based on virtual age models and implemented in the R package VAM**
Presenter: **Laurent Doyen**, Univ Grenoble Alpes, France
Co-authors: Remy Drouilhet
Virtual age models are useful to analyze recurrent events arising in epidemiology (e.g. relapse and treatment times of a disease), industry (e.g. failure and maintenance times of a system), etc. The model consists of a composition of a baseline hazard rate function, characterizing the first time to event distribution, and an effective age function, that allows to take into account events effects. We have developed, and implemented in the R package VAM, a general framework that allows to take into account simultaneously several types of events, having different effects and corresponding to different types of treatments or maintenances. We can also consider planned event times for which the next possible arising time is fixed function of the previous events times and types. The proposed framework is generic in the sense that we propose an iterative way for computing the different characteristics of the model that does not a priori depends on the number of different event types and of their effects. In order to preserve this adaptability in the software implementation, the usage of VAM is based on a formula which specify the characteristics of the data set to analyze and the model used for that. Methods are proposed for events times simulation, maximum likelihood estimation, reliability or health indicators computation. An application to a real data set issued from off-road engines of mining trucks will be presented.

EO170 Room Gordon NON/SEMI PARAMETRIC METHODS FOR STATISTICAL LEARNING Chair: Wei Qian

E1003: **Envelope quantile regression**
Presenter: **Shanshan Ding**, University of Delaware, United States
Co-authors: Zhihua Su, Guangyu Zhu, Lan Wang
Quantile regression offers a valuable complement of classical mean regression for robust and comprehensive data analysis in a variety of applications. We propose a novel envelope quantile regression method (EQR) that adapts a nascent technique called enveloping to improve the efficiency of standard quantile regression. The new method aims to identify material and immaterial information in a quantile regression model and use only the material information for estimation without assuming the quantile regression coefficient vector is sparse. By excluding the immaterial part, the EQR method has the potential to substantially reduce the estimation variability with standard quantile regression. Unlike existing envelop model approaches which mainly rely on the likelihood framework, our proposed estimator is defined through a set of nonsmooth estimating equations. We facilitate the estimation via the generalized method of moments and derive the asymptotic normality of the proposed estimator by applying empirical process techniques. Furthermore, we establish that EQR is asymptotically more efficient than (or at least as asymptotically efficient as) the standard quantile regression estimators without imposing stringent conditions. Hence, the envelope model theory is advanced to general distribution-free settings. We demonstrate the effectiveness of the proposed method via Monte-Carlo simulations and a real data example.

E1170: **An inherent clustering paradigm for supervised and unsupervised learning**
Presenter: **Yiyuan She**, Florida State University, United States
Modern clustering applications are often faced with challenges from high dimensionality and/or nonconvex clusters. The purpose is to give a mathematical formulation of clustering with concurrent dimension reduction and proposes an optimization-based inherent clustering framework. Inherent clustering enjoys a kernel property to work on similarity matrices and can be extended to supervised learning. A simple-to-implement iterative algorithm is developed by use of linearization and block coordinate descent. Nonasymptotic analysis shows the tight error rate of inherent clustering in the supervised setting. Extensive simulations, as well as real-data experiments in network community detection and bioinformatics, demonstrate the excellent performance of the proposed approach.

E1084: **Free-knot splines for generalized linear models**
Presenter: **Jing Wang**, University of Illinois at Chicago, United States
A computational study of bootstrap confidence bands based on a free-knot spline regression is explored for the Generalized Linear Model. In free-knot spline regression, the knot locations as additional parameters offers greater flexibility and the potential to better account for rapid shifts in slope and other important structures in the target function. However, in freeing up the knots, the search for optimal solutions becomes very complicated. In particular, the lethargy property in the objective function results in many local optima with replicate knot solutions. To prevent solutions with identical knots, a penalized Quasi-likelihood estimating equation is proposed that relies on both a Jupp transformation of knot locations and an added penalty on solutions with small minimal distances between knots. Focusing on logistic regression for binary outcome data, a parametric bootstrap is used to study the variability of the proposed estimator and to construct confidence bands for the unknown form of the logistic regression link function. A real example is also studied.

E1341: **On order determination using augmentation predictor**
Presenter: **Wei Luo**, Baruch College, United States
Co-authors: Bing Li
In many statistical dimension reduction problems, including principal component analysis, canonical correlation analysis, independent component analysis, and sufficient dimension reduction, etc., it is often of interest to determine the rank of a matrix parameter based on a consistent matrix estimator. We propose a method called the augmentation estimator for this purpose, with the aid of an augmentation predictor that is artificially generated and merged with the original predictor. The augmentation estimator uses information from both the eigenvalues and the eigenvectors of the matrix estimator. Compared with the existing order-determination methods, it is easy to implement, computationally efficient, consistent under general conditions, and applicable in high-dimensional cases. Its effectiveness is supported by simulation studies. The way we employ the augmentation predictor is novel, which may inspire independent research interest.

EO218 Room Jessel RECENT DEVELOPMENT IN STATISTICAL LEARNING AND MODELING OF COMPLEX DATA Chair: Yanqing Sun

E1486: **A hybrid method for the stratified mark-specific proportional hazards models with missing data with applications**
Presenter: **Yanqing Sun**, University of North Carolina at Charlotte, United States
Co-authors: Li Qi, Peter Gilbert
The motivation comes from the objective to understand how the dengue vaccine efficacy is modified by neutralizing antibodies and whether it depends on dengue genetics. The immune responses in the CYD14 dengue efficacy trial were measured through a two-phase sampling design and there is a high percentage of missing dengue sequences. We develop estimation and hypothesis testing procedures for the stratified mark-specific proportional hazards model with missing covariates and missing marks, where the mark is the genetic distance of an infecting dengue sequence to the dengue sequence represented inside the vaccine. We propose a hybrid method that takes advantages of both the augmented inverse probability weighted method and multiple imputation. A simulation study is conducted to examine the finite-sample performances of the estimators for the mark-specific relative risks and the conditional mark-specific cumulative incidence functions as well as the proposed testing procedures. Our simulation study shows that the proposed hybrid method performs well. The hybrid estimator is better than a direct application of Rubin's multiple imputation with the estimated variances close to the empirical variances, and thus better coverage probabilities. The developed hybrid methods are applied to the CYD14 efficacy trial to assess association of dengue infection with the immune responses.

E1653: **Semiparametric modeling and estimation of the terminal behavior of recurrent marker processes before failure events**
Presenter: **Kwun Chuen Gary Chan**, University of Washington, United States
Co-authors: Mei-Cheng Wang
Recurrent event processes with marker measurements are mostly studied with forward time models starting from an initial event. Interestingly, the processes could exhibit important terminal behavior during a time period before occurrence of the failure event. A natural and direct way to study recurrent events prior to a failure event is to align the processes using the failure event as the time origin and to examine the terminal behavior by a backward time model. We studied regression models for backward recurrent marker processes by counting time backward from the failure event. A three-level semiparametric regression model is proposed for jointly modeling the time to a failure event, the backward recurrent event process, and the marker observed at the time of each backward recurrent event. By jointly modeling the three components, estimating equations can be constructed for marked counting processes to estimate the target parameters in the three-level regression models. The proposed models and methods are illustrated by a community-based AIDS clinical trial to examine opportunistic infections among HIV infected individuals in the last six months of life.

E1356: **Dynamic risk prediction with rank-based survival trees**
Presenter: **Yifei Sun**, Columbia University, United States
Co-authors: Mei-Cheng Wang
Tree-based methods are popular statistical tools for creating simple and interpretable prediction rules with mild assumptions. We introduce a unified framework for tree-structured analysis using binary response or survival data. With arguments from the Neyman-Pearson Lemma, we propose rank-based methods for growing and pruning trees guided by a concordance index. We define the concordance index as a map from an arbitrary function of the covariates to a real number, where the target function maximizes the concordance. In contrast with the existing methods where each split maximizes between node heterogeneity or within node homogeneity, our approach aims to maximize the concordance index of the whole tree. For right-censored survival data, our framework has the flexibility to incorporate time-dependent covariates, resulting in more accurate prognostic models than only considering baseline covariates.

E0563: **Estimation of the optimal surrogate endpoint based on a randomized trial**
Presenter: **Peter Gilbert**, University of Washington & Fred Hutchinson Cancer Research Center, United States
Co-authors: Brenda Price, Mark van der Laan
A common scientific problem is to determine a surrogate outcome for a long-term outcome so that future randomized studies can restrict themselves to only collecting the surrogate outcome. We consider the setting that we observe n independent and identically distributed observations of a random variable consisting of baseline covariates, a treatment, a vector of candidate surrogate outcomes at the intermediate time point, and the final outcome of interest at a final time point. It is assumed that in this current study the treatment is randomized, conditional on the baseline covariates. The goal is to use these data to learn a most-promising surrogate for use in future trials for estimation and testing of a mean contrast treatment effect on the outcome of interest. We define an optimal surrogate for the current study as the function of the data collected by the intermediate time point that satisfies the Prentice definition of a valid surrogate endpoint and that optimally predicts the final outcome (in the current study): this optimal surrogate is a function of the data generating distribution and is thus unknown. We show that this optimal surrogate is a conditional mean and present super-learner and targeted super-learner based estimators that can accommodate high-dimensional covariates and response endpoints,whose predicted outcomes are used as the surrogate in applications.

E1438: **A surrogate-based approach to modelling the impact of hydrodynamic shear stress on biofilm deformation**
Presenter: **Oluwole Oyebamiji**, Newcastle University, United Kingdom
Co-authors: Darren Wilkinson
The aim is to investigate the feasibility of using a surrogate-based method to emulate the deformation and detachment behaviour of a biofilm in response to hydrodynamic shear stress. The influence of shear force and growth rate parameters on the patterns of growth, structure and resulting shape of microbial biofilms was examined. We develop a novel statistical modelling approach to this problem, using a combination of Bayesian Poisson regression and dynamic linear models for the emulation. We observe that the hydrodynamic shear force affects biofilm deformation in line with some literature. Sensitivity results also showed that the shear flow and yield coefficient for heterotrophic bacteria are the two principal mechanisms governing the bacteria detachment. The sensitivity of the model parameters is temporally dynamic, emphasising the significance of conducting the sensitivity analysis across multiple time points. The surrogate models are shown to perform well, and produced 480 fold increase in computational efficiency. We conclude that a surrogate-based approach is effective, and resulting biofilm structure is determined primarily by a balance between bacteria growth and applied shear stress.

EO653 Room Montague ANALYSIS OF INCOMPLETE DATA Chair: Shaun Seaman

E0212: **Relative efficiency of joint-model and full-conditional-specification multiple imputation when models are compatible**
Presenter: **Shaun Seaman**, University of Cambridge, United Kingdom
Co-authors: Rachael Hughes
Fitting a regression model of interest is often complicated by missing data on the variables in that model. Multiple imputation (MI) is commonly used to handle these missing data. Two popular methods of MI are joint model MI and full-conditional-specification (FCS) MI. These are known to yield imputed data with the same asymptotic distribution when the conditional models of FCS are compatible with the joint model. We show that this asymptotic equivalence of imputation distributions does not imply that joint model MI and FCS MI will also yield asymptotically equally efficient inference about the parameters of the model of interest, nor that they will be equally robust to misspecification of the joint model. When the conditional models used by FCS MI are linear, logistic and multinomial regressions, these are compatible with a restricted general location (RGL) joint model. We show that MI using the RGL joint model (RGL MI) can be substantially more asymptotically efficient than FCS MI, but this typically requires very strong associations between variables. When associations are weaker, the efficiency gain is small. Moreover, FCS MI is shown to be potentially much more robust than RGL MI to misspecification of the RGL model when there is substantial missingness in the outcome variable.

E0276: **Multiple imputation and multivariable model building**
Presenter: **Tim Morris**, MRC Clinical Trials Unit at UCL, United Kingdom
Data arising from randomised trials and cohort studies are often used to build prognostic models. However, incomplete covariate data are often present in these datasets. It was only in the mid-2000s that authors began to attempt to build prognostic models using multiple imputation, which has since become more common. However, there are many barriers to being able to achieve the sophistication of model building in complete data when using multiple imputation. We will review some of the issues in combining model building with multiple imputation. These include: specifying the imputation model; allowing departures from 'missing at random'; testing variables for inclusion/exclusion; accommodating non-linear effects; and finally, for survival data, permitting time-varying effects. Issues in the imputation phase revolve around being able to specify a model that is at least as rich as the final selected model. Issues in the model building phase arise due to the lack of a meaningful likelihood in multiply-imputed data. The assessment of model performance is again more complex following multiple imputation. we will describe recent and current work on these challenges and some future directions.

E1212: **Logistic regression with missing continuous and categorical data**
Presenter: **Wei Jiang**, Ecole Polytechnique, France
Co-authors: Julie Josse, Erwan Scornet
To make inference with missing values, a recommended approach consists in using an EM algorithm to obtain maximum likelihood estimates and a supplemented EM algorithm for the variance. However, it is often said that it is not necessary straightforward to derive such algorithms and it can be observed that indeed this approach is almost never used in practice neither implemented. The use of (multiple) imputation is more popular and has the great advantage that it is not only designed for one statistical method but it allows to carry many analyses from the same data. We will thoughtfully compare both approaches to perform logistic regression with missing values and both categorical and continuous data. We will present a quite straightforward Monte Carlo EM based on Sampling Importance Resampling which can be used as an alternative to the computationally adaptive rejection sampling within Gibbs in the framework of GLM. Imputation methods include recent proposal for mixed data based on principal component methods and non-parametric Bayesian. The methods will be illustrated on the analysis of a large register from the Paris Hospital (APHP) to model the decisions and events when severe trauma patients are handled by emergency doctors.

E1144: **Propensity score analysis with partially observed confounders: Multiple imputation and the missingness pattern approach**
Presenter: **Clemence Leyrat**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: James Carpenter, Elizabeth Williamson
Propensity scores (PS) estimate the probability of an individual being treated given their characteristics. They are commonly used to address confounding bias in observational studies. One popular method to achieve covariate balance between exposure groups is to re-weight individuals by the inverse of their PS value (Inverse Probability of Treatment Weighting). In applications, a major issue is how to estimate the PS when confounders are partially observed. Multiple imputation (MI) is a natural, and widely used tool. However, in the PS context a number of key issues remain unresolved, included (i) how to build the imputation model to ensure compatibility between the substantial, the imputation and the PS models, (ii) how to apply Rubins rules in the PS context, and (iii) how to extend MI to address the question of time-varying treatment and exposure. An alternative strategy, the missingness pattern approach (MPA), has been proposed: although simpler to implement than MI, the assumptions required for its used are unclear, so we investigate this question further. We explore the performance of MI and MPA both theoretically and using a simulation study. We thus provide guidelines to perform a PS analysis in presence of missing data.

E1108: **The not at random fully conditional specification procedure**
Presenter: **Daniel Tompsett**, MRC Biostatistics Unit:University of Cambridge, United Kingdom
The Not at Random Fully Conditional Specification Procedure (NARFCS) for imputing multivariate missing data under the Missing Not at Random (MNAR) assumption will be described. Methods exist to impute MNAR data under the Fully Conditional Specification procedure (FCS). The NARFCS procedure represents a formalisation of such methods, which fully defines the imputation models for each variable with missing data. We will first outline the imputation procedure, and offer general advice as to how to construct it's imputation models. We will then show how to elicit values for the sensitivity parameters of the procedure, which are conditional on the remaining variables of the data. The procedure will be demonstrated on an example dataset using functions for R, specifically designed for NARFCS.

EO455 Room Senate RECENT DEVELOPMENTS ON PROGRAM EVALUATION METHODS Chair: Sebastian Calonico

E1197: **Regression discontinuity with many thresholds**
Presenter: **Marinho Bertanha**, University of Notre Dame, United States
Numerous empirical studies employ regression discontinuity designs with multiple cutoffs and heterogeneous treatments. A common practice is to normalize all the cutoffs to zero and estimate one effect. This procedure identifies the average treatment effect (ATE) on the observed distribution of individuals local to existing cutoffs. However, researchers often want to make inferences on more meaningful ATEs computed over general counterfactual distributions of individuals rather than simply the observed distribution of individuals local to existing cutoffs. A consistent and asymptotically normal estimator for such ATEs is proposed when heterogeneity follows a non-parametric function of cutoff characteristics in the sharp case. The proposed estimator converges at the minimax optimal rate of root-n. Identification in the fuzzy case with multiple cutoffs is impossible unless heterogeneity follows a finite dimensional function of cutoff characteristics. Under parametric heterogeneity, an ATE estimator for the fuzzy case is proposed that optimally combines observations to maximize its precision.

E1229: **Multivalued treatments and decomposition analysis: An application to the WIA program**
Presenter: **Sebastian Calonico**, University of Miami, United States
Co-authors: Ying-Ying Lee, Wallice Ao
Efficient estimators are analyzed for multi-valued treatment effects on the treated that can be used to conduct distributional decompositions in the outcome variable. In particular, we propose two-step semiparametric efficient estimators that can be used to decompose differences in the outcome distribution into (i) a wage structure effects, arising due to the conditional outcome distributions associated with different levels of participation; and (ii) a composition effect, arising due to differences in the distribution of observable characteristics. These counterfactual differences reveal causal relationships under a conditional independence assumption. Moreover, we calculate the semiparametric efficiency bound for the multivalued treatment effects and we provide uniform inference results for efficient multi-valued nonparametric propensity score weighting estimators.We employ these procedures to study evaluate the Workforce Investment Act (WIA), a large US job service program. Our estimation results show that heterogeneity in levels of participation is an important dimension to evaluate the WIA and other social programs in which participation varies. The results, both theoretically and empirically, provide rigorous assessment of intervention programs and relevant suggestions to improve the performance and cost-effectiveness of these programs.

E1117: **Two-step estimation and inference with possibly many included covariates**
Presenter: **Xinwei Ma**, University of Michigan, United States
Co-authors: Matias Cattaneo, Michael Jansson
The implications of including many covariates in a first-step estimate entering a two-step estimation procedure are studied. We find that a first order bias emerges when the number of covariates is "large" relative to the sample size, rendering standard inference procedures invalid. We show that the jackknife is able to estimate this "many covariates" bias consistently, thereby delivering a new fully automatic bias-corrected two-step point estimator. The jackknife also consistently estimates the standard error of the original two-step point estimator (prior jackknife bias-correction). For inference, we develop a valid post-bias-correction bootstrap approximation that accounts for the additional variability introduced by the jackknife bias-correction. We find that the jackknife bias-corrected point estimator and the bootstrap post-bias-correction inference perform excellent in simulations, offering important improvements over conventional two-step point estimators and inference procedures, which are not robust to including many covariates. We apply our results to an array of distinct treatment effect and policy evaluation settings. In particular, we discuss in detail Marginal Treatment Effect (MTE) and Local Average Response Function (LARF) estimation in instrumental variables settings: our results are the first to offer valid estimation and inference when many instruments/covariates are included in non-linear settings with heterogeneous treatment effects.

E1097: **Schooling and training for unemployed workers**
Presenter: **Zhuan Pei**, Cornell University, United States
Co-authors: Pauline Leung
Recognizing that the labor market increasingly values highly skilled workers, recent U.S. policies have focused on encouraging the unemployed to pursue further education. However, while the returns to education for traditional students is unambiguously large, the evidence for unemployed workers is scarce. We use linked administrative data to characterize the type of unemployed workers who seek additional education during and in the aftermath of the Great Recession. In addition, we measure the effect of additional education on future labor market outcomes using various matching techniques.

EO035 Room Woburn STATISTICAL NETWORK ANALYSIS Chair: Chenlei Leng

E0857: **Community detection with covariates**
Presenter: **Yang Feng**, Columbia University, United States
Community detection is one of the fundamental problems in the study of network data. Most existing community detection approaches only consider edge information as inputs, and the output could be suboptimal when covariates are available. In such cases, it is desirable to leverage covariates information for the improvement of community detection accuracy. Towards this goal, we propose flexible network models incorporating covariates, and develop likelihood-based inference methods. For the proposed methods, we establish favorable asymptotic properties as well as efficient algorithms for computation. Numerical experiments show the effectiveness of our methods in utilizing covariates across a variety of simulated and real network data sets.

E0714: **Fixed-effect regressions on network data**
Presenter: **Koen Jochmans**, Sciences Po, France
Co-authors: Martin Weidner
Inference on fixed effects in a linear regression model estimated from network data is studied. An important special case of our setup is the two-way regression model, which is a workhorse method in the analysis of matched data sets. Networks are typically quite sparse and it is difficult to see how the data carry information about certain parameters. We derive bounds on the variance of the fixed-effect estimator that uncover the importance of the structure of the network. These bounds depend on the smallest non-zero eigenvalue of the (normalized) Laplacian of the network and on the degree structure of the network. The Laplacian is a matrix that describes the network and its smallest non-zero eigenvalue is a measure of connectivity, with smaller values indicating less-connected networks. These bounds yield conditions for consistent estimation and convergence rates, and allow to evaluate the accuracy of first-order approximations to the variance of the fixed-effect estimator. The bounds are also used to assess the bias and variance of estimators of moments of the fixed effects.

E0719: **Consistent maximum likelihood estimation of random graph models with local dependence and growing neighborhoods**
Presenter: **Michael Schweinberger**, Department of Statistics, Rice University, United States
Co-authors: Jonathan Stewart
In general, statistical inference for exponential-family random graph models of dependent random graphs given a single observation of a random graph is problematic. We show that statistical inference for exponential-family random graph models holds promise as long as models are endowed with a suitable form of additional structure. We consider a simple and common form of additional structure called multilevel structure. To demonstrate that exponential-family random graph models with multilevel structure are amenable to statistical inference, we develop the first concentration and consistency results covering maximum likelihood estimators of a wide range of full and non-full, curved exponential-family random graph models with local dependence and natural parameter vectors of increasing dimension. In addition, we show that multilevel structure facilitates local computing of maximum likelihood estimators and in doing so reduces computing time. Taken together, these results suggest that exponential-family random graph models with multilevel structure constitute a promising direction of statistical network analysis.

E1785: **A random effects stochastic block model for community detection in multiple networks with applications to neuroimaging**
Presenter: **Subhadeep Paul**, The Ohio State University, United States
Co-authors: Yuguo Chen
Motivated by multi-subject and multi-trial experiments in neuroimaging studies, we develop a modeling framework for joint community detection

in a group of related networks, which can be considered as a sample from a population of networks. The proposed model, which we call the random effects stochastic block model, facilitates the study of group differences and subject specific variations in the community structure. The model proposes the existence of a putative mean community structure which is representative of the group or 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 or trials. We propose two methods to estimate the quantities of interest, a variational EM algorithm and two model-free "two-step" methods based on spectral and non-negative matrix factorization respectively. We also develop several resampling based hypothesis tests to test for differences in community structure in two populations of subjects both at the whole network level and node level. The methodology is applied to publicly available fMRI datasets from multi-subject experiments involving schizophrenia patients along with healthy controls.

E1697: **Community structures in (socially) connected systems**
Presenter: **Roy Cerqueti**, University of Macerata, Italy

The aim is to elaborate on the community structures of complex networks whose links are also of indirect and social type. At this aim, an extension of the clustering coefficients of the nodes is introduced and discussed. The usefulness of the assessment of the communities in a network for describing its riskiness is explored with peculiar attention. Some remarks on the discrepancies between social resilience and classical resilience are also carried out. In so doing, a definition of risk of general socio-economic nature is provided. Paradigmatic examples illustrate the theoretical results and validate the framework, assisting also the reader in having an intuitive view of the presented concepts.

EO390 Room SH349 ADVANCED STATISTICS FOR UNDERSTANDING THE EVOLUTION OF CANCER Chair: Inigo Martincorena

E1235: **Statistical approaches to unravel the life history of cancers**
Presenter: **Stefan Dentro**, Wellcome Trust Sanger Institute, United Kingdom
Co-authors: David Wedge, Peter Van Loo

Tumours evolve through the gradual acquisition of somatic mutations in the DNA of their cells. Some of these mutations provide a selective advantage to cell in which it occurred, giving rise to clonal expansions that allow the tumour to evolve and adapt. To better understand this process we have developed algorithms to disentangle the life history of tumours from a single or from multiple biopsy massively parallel sequencing data. These methods perform statistical inference over mutations and their properties, which are used as markers of tumour subpopulations that are in the process of expansion. Our algorithms are applied, as part of a larger effort, to the 2778 whole genome sequences in the International Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes project. We observe that nearly all tumours contain at least one distinct subpopulation, that the subpopulations contain mutations in known cancer genes and that these mutations are under positive selection. Furthermore, in the subpopulations we find mutations in genes relevant in the clinic. These findings suggest tumours continue to evolve up until diagnosis and may inform treatment choices.

E1285: **Reliable detection of selection in cancer**
Presenter: **Inigo Martincorena**, Wellcome Trust Sanger Institute, United Kingdom

Cancer is largely caused by somatic mutation and clonal selection. Yet, despite the central role of selection in cancer, tools to accurately quantify selection in cancer genomes are lacking. We will refine statistical methods from the field of molecular evolution to study selection in cancer. Applying these methods to 8,000 cancer exomes reveals that somatic evolution is characterised by a distinct and universal pattern of selection never seen in species evolution and clarifies the extent of positive and negative selection in cancer genomes.

E1147: **PhyClone: A forest structured Chinese restaurant process for inferring tumour phylogenies**
Presenter: **Andrew Roth**, University of Oxford, United Kingdom
Co-authors: Alexandre Bouchard

Cancer is a disease caused by the ongoing accumulation of genomic alterations. Once a cancer has formed this process generates heterogeneity within the cancer cell population. We consider the problem of inferring the phylogenetic tree relating cancer cell populations from patient tumours using bulk genome sequencing data. Unlike data in traditional phylogenetic problems, bulk sequencing data measures an admixture of taxa in the tree. Furthermore, taxa from all nodes in the tree, not just leafs, may be represented. In previous work we have shown how clustering mutations which occur at the same cellular prevalence can be used to identify cancer cell populations. We extend that approach to also infer the underlying tree relationship among populations. We develop a non-parametric Bayesian prior over the clustering of the data and tree structure relating clusters. Posterior inference using Markov chain Monte Carlo (MCMC) sampling is performed using an auxiliary variable scheme and conditional Sequential Monte Carlo (cSMC) sampling. One advantage of this model is that it is possible to marginalize the parameters associated with the nodes of the trees which represent the cellular prevalence of mutations. Collapsing these variables can potentially improve the mixing of the MCMC sampler. We compare this model to previous work which used the tree structured stick breaking process, another non-parametric Bayesian prior over tree topologies and clusterings.

E1819: **Reconstructing cancer phylogenies with a small number of pairwise haplotypes and bulk tumour sequencing data**
Presenter: **Amit Deshwar**, University of Toronto, Canada

PhyloSpan, a novel method for reconstructing the evolutionary history of tumours from sequencing data from one or more heterogeneous tumour samples, is presented. PhyloSpan uses data from reads (or read-pairs) that cover the genomic loci of more than one somatic mutation – in addition to SNV allele frequencies and CNA prevalence estimates used by other methods. We show that these haplotypes can resolve ambiguities in phylogenetic reconstruction and can increase the sensitivity of reconstruction methods to subclones with similar cellular prevalences. We then considered the data from 2,778 PCAWG tumour samples. By phasing SNV pairs in haploid regions in the 775 PCAWG samples with appropriate pairs, we estimate that less than 4% of SNV pairs, on average, are in branching lineages, strongly suggesting that the vast majority of subclones are in linear phylogenies consistent with incomplete selective sweeps. Applying PhyloSpan to PCAWG samples with pairwise haplotype data, we find numerous cases where PhyloSpan can distinguish subclones with similar cellular prevalences; can merge subclonal clusters that are incorrectly deemed separate; and can detect branching lineages undetectable with standard techniques. PhyloSpan can readily applied to long read sequencing data; but even short read sequencing data often contains informative haplotypes than can be leveraged to improve cancer phylogenies.

E1722: **Modelling the early menarche and late menopause in breast cancer screening through CGAMLSS models**
Presenter: **Elisa Duarte**, Universidade de Santiago de Compostela, Spain
Co-authors: Carmen Cadarso Suarez, Bruno de Sousa, Giampiero Marra, Rosalba Radice, Vitor Rodrigues

It is a fact that in the breast cancer etiology the risk of the disease increases with reproductive factors as early menarche and late menopause. It is believed that longer is reproductive lifespan (difference between menopause and menarche ages), higher is the risk of breast cancer. This is such an important issue that there are several studies addressing the trend of menarche and menopause ages along the years. The data provided by the Portuguese Cancer League, Central Branch sponsored by the Breast Cancer Screening Program in 78 municipalities refers to the breast cancer screening program over a period of twenty years. The dataset has information about 212,517 postmenopausal women, born between 1920 and 1965 who have attended the breast cancer screening program in central region of Portugal. For this analysis we employ Bivariate Copula Additive

Models for Location, Scale and Shape considering the menopause and menarche ages as binary outcomes. The CGAMLSS models extend the scope of univariate GAMLSS using a copula approach. They allow modeling the marginal distributions of different families and types, and the dependence structure using predictor equations that can be flexibly specified using smoothers with single or multiples penalties, hence allowing for several types of covariate effects. All the models parameters are estimated simultaneously. The inference is carried out using the R package GJRM.

EG020 Room G21A CONTRIBUTIONS IN VARIABLE SELECTION Chair: Jan Gertheiss

E1476: **Detection of influential points as a byproduct of resampling-based variable selection procedures**
Presenter: **Riccardo De Bin**, University of Oslo, Norway
Co-authors: Anne-Laure Boulesteix, Willi Sauerbrei

Influential points can cause severe problems when deriving a multivariable regression model. A novel approach to check for such points is proposed, based on the variable inclusion matrix, a simple way to summarize results from resampling-based variable selection procedures. The variable inclusion matrix reports whether a variable is included in a regression model fitted on a pseudo-sample generated from the original data. It is used to study the variable selection stability, to derive weights for model averaged predictors and in others investigations. Concentrating on variable selection, it also allows understanding whether the presence of a specific observation has an influence on the selection of a variable. From the variable inclusion matrix, indeed, the inclusion frequency (I-frequency) of each variable can be computed only in the pseudo-samples which contain the specific observation. When the procedure is repeated for each observation, it is possible to check for influential points through the distribution of the I-frequencies, visualized in a boxplot, or through a Grubbs' test. Outlying values in the former case and significant results in the latter point to observations having an influence on the selection of a specific variable and therefore on the finally selected model. This novel approach is illustrated in two real data examples.

E1840: **Combinatorial strategies for greedy regression selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Georgiana-Elena Pascaru, Erricos John Kontoghiorghes

Greedy kind algorithms are an established approach to the regression model selection problem. A main drawback of these methods is the reduced number of submodels that are evaluated in order to select a solution, thus failing to find the optimum. Two strategies that aim to overcome this issue are investigated: selection–k (SEL–k) and tree selection–k (TREE–k). SEL–k builds on standard forward selection, but selects at each step the best k variables, instead of the only best one, thus allowing a degree of correlation between the variables included in the solution model. TREE–k is a method that explores a combinatorial search space, thus increasing the number of submodels that are investigated. Specifically, at each step of the algorithm a new search brunch is considered for each of the best k most significant variable, say $x_1, \cdots, x_k$. On each of the i–th branch ($i = 1, \cdots, k$), the selection will always include x_i in the submodel. A branch will terminate either when there are no more significant variables to choose from, or when all variables have been considered. Various experiments are conducted on both real and artificially generated datasets in order to assess the two proposed algorithms. The results are presented and discussed.

E0299: **Variable selection in quantile varying coefficient models with heteroscedastic error**
Presenter: **Mohammed Abdulkerim Ibrahim**, Hasselt University, Belgium
Co-authors: Anneleen Verhasselt

Quantile regression gives a thorough view of the relationship of the covariates with the entire distribution of the response. Varying coefficient models are considered, allowing the coefficients vary with time in a longitudinal data setting. Since important variables can influence various quantiles in different ways, the problem of variable selection in quantile regression is more challenging. We propose an easy way to check the influence of the covariates on the distribution of the response by investigating both the location and the scale. The functions are estimated with Penalized B-splines. A grouped Adaptive Lasso and nonnegative garrote are considered for variable selection. The selection procedures have consistency in variable selection under suitable conditions. Further, the estimated functional coefficients are shown to have an optimal convergence rate to the true functional coefficients. Our simulation study shows that both methods have good performance with respect to both the location and the scale in selecting the important covariates as well as in estimating the functional coefficients. The procedures are compared with grouped adaptive Lasso using B-splines estimation and group SCAD. Nonnegative garrote outperforms the other methods way far with respect to the computational time. Finally, the procedures are illustrated on two real-data examples.

E1654: **Regularization parameter selection for sparse methods via AIC**
Presenter: **Yoshiyuki Ninomiya**, Kyushu University, Japan

Generally sparse methods contain a regularization parameter that determines the result, and several information criteria have been proposed for its selection. While any of them would assure consistency in model selection, we have no appropriate rule to choose between the different possible criteria. On the other hand, a finite correction to the AIC has been provided in a Gaussian linear regression setting for the LASSO. The finite correction is theoretically assured from the viewpoint not of the consistency but of minimizing the prediction error, and it does not have the above-mentioned difficulty in the choice. In general, however, the finite correction cannot be obtained in the case of generalized linear models, and so we derive a criterion from the original definition of the AIC, that is, an asymptotically unbiased estimator of the Kullback-Leibler divergence. Our criterion can be easily obtained and requires fewer computational tasks than does cross-validation, but its performance is almost the same as or superior to that of cross-validation.

E1631: **Principal component regression for generalized linear models via L1-type regularization**
Presenter: **Shuichi Kawano**, The University of Electro-Communications, Japan
Co-authors: Hironori Fujisawa, Toyoyuki Takada, Toshihiko Shiroishi

Principal component regression (PCR) is a widely used two-stage procedure: principal component analysis (PCA), followed by regression in which the selected principal components are regarded as new explanatory variables in the model. Note that PCA is based only on the explanatory variables, so the principal components are not selected using the information on the response variable. To address this problem, we propose a one-stage procedure for PCR in the framework of generalized linear models. The loss function is based on a combination of the regression loss and PCA loss. An estimate of the regression parameter is obtained as the minimizer of the loss function with an L1-type regularization term. We call this method sparse principal component regression for generalized linear models (SPCR-glm). SPCR-glm enables us to obtain sparse principal component loadings that are related to a response variable. Numerical results are given to illustrate the effectiveness of SPCR-glm.

EG014 Room G5 CONTRIBUTIONS IN HMM AND MCMC Chair: Robert Kohn

E1652: **Bayesian analysis of predictive non-Homogeneous hidden Markov models using Polya-Gamma data augmentation**
Presenter: **Constandina Koki**, Athens University of Economics and Business, Greece
Co-authors: Ioannis Vrontos, Loukia Meligkotsidou

Non-Homogeneous hidden Markov Models (NHMMs) are considered for forecasting univariate time series. We introduce two state NHMMs where the time series are modeled via different predictive regression models for each state. Furthermore, the time-varying transition probabilities depend on exogenous variables through a logistic function. In a hidden Markov setting, inference for logistic regression coefficients becomes complicated and in some cases impossible due to convergence issues. To address this problem, we use a new latent variable scheme, that utilizes the Polya-Gamma class of distributions. Given an available set of predictors, we allow for model uncertainty regarding the predictors that affect the series both linearly – in the mean – and non-linearly – in the transition matrix. Variable selection and inference on the model parameters are based on a MCMC scheme with reversible jump steps. Single-step and multiple-steps-ahead predictions are obtained based on the most probable model, median probability model or a Bayesian Model Averaging (BMA) approach. Simulation experiments, including an empirical study on real financial data, illustrate the performance of our algorithm in various setups, in terms of mixing properties, model selection and predictive ability.

E0853: **Rejection-free ensemble MCMC with applications to factorial hidden Markov models**
Presenter: **Kaspar Martens**, University of Oxford, United Kingdom
Co-authors: Michalis Titsias, Christopher Yau

Bayesian inference for complex models is challenging due to the need to explore high-dimensional spaces in the presence of multimodality. Standard Monte Carlo samplers can have difficulties effectively exploring the posterior landscape and are often restricted to exploration around localised regions that depend on initialisation. We introduce a general purpose rejection-free ensemble Markov Chain Monte Carlo (MCMC) technique to improve on existing poorly mixing samplers. This is achieved by combining parallel tempering and an auxiliary variable move to exchange information between the chains. We demonstrate this ensemble MCMC scheme on Bayesian inference in Factorial Hidden Markov Models. This high-dimensional inference problem is difficult due to the exponentially sized latent variable space. Existing sampling approaches mix slowly and become trapped in local modes. We show that the performance of these samplers is improved by our rejection-free ensemble technique and that the method is attractive and "easy-to-use" since no parameter tuning is required.

E1439: **Estimation of a Poisson autoregressive hidden Markov process with Poisson regression-type measurement errors**
Presenter: **Ruzzel Ragas**, University of the Philippines-Diliman, Philippines

A method of estimation for Poisson autoregressive model when data is contaminated with error is proposed in the context of hidden Markov modeling paradigm. Model parameters are then estimated by its maximum likelihood estimator computed using data cloning method. Subsequently, particle filter is used to estimate the hidden process. Simulation studies indicate that bias and standard error of the parameter estimates decreases as length of the time series increases. However, misspecification of the covariate parameter leads to poor predictive ability of the model, while in most cases, predictive ability is not affected by the length of the time series.

E1748: **Markov chains models with time-varying parameters**
Presenter: **Lionel Truquet**, ENSAI, France

The focus is on time-inhomogeneous Markov chains on general state spaces and for which the finite-dimensional distributions can be approximated locally by ergodic Markov chains via an infill asymptotic. Our approach, which is based on contraction properties of Markov kernels, can be seen as a Markov version of the notion of local stationarity. In particular, one can consider time-inhomogeneous autoregressive processes, finite-state Markov chains or Markov-switching processes for which statistical inference is still possible. We will explain how to construct such models and what type of statistical results it is possible to get.

E1549: **Evolutionary computation and multiple chains MCMC sampling: An overview**
Presenter: **Manuel Rizzo**, Sapienza University of Rome, Italy

Evolutionary Computation (EC) has been introduced in the 1960s in the field of Artificial Intelligence, and includes a variety of algorithms (called Evolutionary Algorithms, EAs) designed to analyze complex systems and problems. These tasks are accomplished by use of metaphors related to evolutionary biology and Darwins principles, for which a population of interacting individuals evolves through generations to populations better able to adapt the environment. In the last twenty years EAs have been employed to a wide range of statistical problems; a particular application is Markov Chain Monte Carlo (MCMC) sampling with multiple chains running in parallel. This framework, motivated by situations in which target distribution exhibits multimodality, high-dimensionality or has highly correlated components, suggests many analogies with EAs if chains are allowed to interact with each other. This has led to several contributions and algorithms in literature by researchers from different fields of science. We aim at reviewing these different proposals, identifying peculiar procedures to deal with MCMC issues, for example preservation of convergence to desired equilibrium distribution, highlighting strengths and weaknesses, and unifying them in a common framework of EC.

CO374 Room MAL B18 NEW DEVELOPMENT ON NONLINEAR TIME SERIES AND ITS APPLICATIONS Chair: Jia Chen

C0227: **Simultaneous specification testing for nonlinear time series models**
Presenter: **Shuo Li**, Tianjin University of Finance and Economics, China
Co-authors: Bin Guo, Yundong Tu

A simultaneous test is proposed for the specification of the conditional mean and conditional variance functions as well as the error distribution in nonlinear time series models. Constructed by comparing two different density estimators for the response variable, the proposed test has a Gumbel limiting distribution under the null hypothesis and is consistent against a general class of alternative hypotheses. A parametric bootstrap procedure is proposed to approximate its finite sample distribution. The proposed test is shown to have nice performances in extensive simulations. The application to the continuous time diffusion model is illustrated via an analysis on the U.S. Federal fund rate data.

C0309: **Factor-augmented time series models with functional coefficients**
Presenter: **Jiraroj Tosasukul**, The University of York, United Kingdom

A new class of functional-coefficient time series models is introduced, where the regressors consist of autoregressors and latent factor regressors, and the coefficients vary with the certain index variable. The unobservable factor regressors are estimated through imposing an approximate factor model on very high dimensional exogenous time series variables and subsequently implementing the classical principal component analysis. With the estimated factor regressors, a local linear smoothing method is used to estimate the coefficient functions and obtain a one-step ahead nonlinear forecast of the response variable, and then a bootstrap procedure is introduced to construct the prediction interval. In particular, our asymptotic theory shows that the local linear estimator and the nonlinear forecast using the estimated factor regressors are asymptotically equivalent to those using the true latent factor regressors. The developed methodology is further extended to the case of multivariate response vectors and the model is generalised to the factor-augmented vector time series model with functional coefficients. The latter substantially generalises the linear factor-

augmented vector autoregressive model which has been extensively studied in the literature. Finally, some simulation studies and an empirical application are given to examine the finite-sample performance of the proposed models and methodologies.

E1843: **Dynamic semiparametric factor model with structural breaks**
Presenter: **Weining Wang**, City U of London, United Kingdom
Co-authors: Wei Biao Wu, Likai Chen
A dynamic semiparametric model with structural factor dynamics for structral breaks is considered. The observations are described by a few low dimensional factors with time invariable functions of covariates as loadings. This framework could incorporate both discontinuity (structural breaks)in space and in time. The discontinuity in space is to account for possible known spatial discrepancies, while the unknown structural break in time models the regime switching effects introduced by exogenous shocks over time. We provide identifiability condition for our factor processes with structural breaks. We develop an estimation procedure allowing to estimate and make inference on the break point in time. The procedure's precision is evaluated via a simulation study. Finally we show two empirical illustrations on modeling the dynamics of the minimum wage in China and the dynamics of trading volume from a limit order book dataset.

C0690: **Enhanced Sharpe ratio via eigen portfolios selection**
Presenter: **Chengguo Weng**, University of Waterloo, Canada
The aim is to show how to pick optimal portfolios by modulating the impact of estimation risk in the covariance matrix. The portfolios are selected to maximize their Sharperatios. Each eigenvector of the covariance matrix corresponds to a maximum Sharpe ratio(MSR) portfolio for a different set of expected returns. Assuming the portfolio manager has views on the future expected returns, a portfolio consistent with her views can be approximated by the first K eigenvectors of the covariance matrix. Since the estimation error in the covariance matrix tends to be most severe in the eigenvectors associated with the smallest eigenvalues, the elimination of the tail eigenvectors reduces estimation error. We substitute the vector of expected excess returns by its lower-dimensional approximation, so that the MSR portfolio is not contaminated by the estimation errors in the tail. We show the equivalence between the expected returns approximation approach and the spectral cut-off method of regularizing a precision matrix. We introduce a more general spectral selection method, which uses non-consecutive eigenvectors to approximate the expected excess returns. The spectral methods, when applied to empirical data yield Sharpe ratios consistently higher than those of equally weighted portfolios.

C1849: **State space approach to online learning and forecasting in mixture autoregressive model**
Presenter: **Mohammadreza Yeganegi**, Islamic Azad University, Central Tehran Branch, Iran
Co-authors: Rahim Chinipardaz
Online learning (parameter estimation) and forecasting of time series models are of great interest in both dynamic control and forecasting financial markets. The focus is on online parameter estimation in mixture autoregressive time series using state space approach. The state space representation of mixture autoregressive is given and stability and steady state of this representation is investigated using simulation study. The EM and online EM algorithms is organized for parameter estimation in state space. The performance of proposed method is also investigated using simulation study.

CO611 Room MAL 414 QUANTITATIVE AND STATISTICAL METHODS IN FINANCE Chair: Jan Vecer

C1287: **A factor-model approach to correlation stress testing**
Presenter: **Natalie Packham**, Berlin School of Economics and Law, Germany
Co-authors: Fabian Woebbeking
A factor-model is developed for parameterising correlation matrices of financial portfolios. The factor-model structure allows to understand various drivers of correlation amongst portfolio constituents. The approach can be used to translate economic scenarios into constraints and changes on the dependence structure allowing to measure the impact of specific scenarios on portfolio risk. In a detailed case study, we apply the factor-model approach to the credit derivatives trading strategy by JP Morgan Chase, dubbed the London Whale, that led to losses in the magnitude of 6.2 bln USD in 2012. By choosing the factors according to the stated investment goals of the strategy (e.g. hedging investment grade credit derivatives with high-yield credit derivatives), we derive insights into the specific correlation risks inherent in the trading strategy.

C1502: **Hierarchical clustering of equities with the Fischer information metric**
Presenter: **Stephen Taylor**, New Jersey Institute of Technology, United States
Information Geometry offers a correspondence between differential geometry and statistics through the Fisher Information matrix. In particular, given two models from the same parametric family of distributions, one can compute the distance between these models using only their parameters and the Fisher Information matrix for this family. One practical limitation of this distance is that it is often difficult to calculate. We review such complications and provide a general form for the distance function for one parameter models. We next focus on higher dimensional extreme value models including the Pareto distribution and discuss how to use shooting point methods to solve the geodesic equation. Finally, we present an application where we first fit extreme value distributions using maximum likelihood estimation to all S&P 500 stocks and then compute the pairwise distances between their best fit parameters. This is used as an input into a hierarchical clustering algorithm to provide a tail risk based clustering of the securities.

C1803: **Performance of volatility maximization strategies**
Presenter: **Jan Vecer**, Charles University, Czech Republic
Volatility in finance is traditionally regarded negatively. The classical risk measures are increasing functions of volatility. Thus a traditional portfolio management tries to completely eliminate or minimize the risk associated with volatility. However, volatility is just a measure of dispersion, so a high volatility can result in both substantially negative or substantially positive outcomes. We show that the volatility maximization portfolio is maximizing the costs associated with insuring an actively traded portfolio. These contracts are called options on a traded account and their existence in practice is very limited precisely for the high price. Option pricing theory uses risk neutral measures (also called martingale measures) for pricing in contrast to using the real measure. The martingale measures determine the replication costs of the scenarios insured by the option. These replication costs can differ substantially from the real expectation and thus the risk neutral measures and the real measures may exhibit some discrepancies. In fact, they differ most on the scenarios associated with high volatility. In particular, the substantially negative outcomes for high volatility strategies are much less likely to happen in reality in comparison to their replicating costs. We illustrate the performance of the high volatility strategies on the portfolio of major world currencies and a portfolio of stocks from NASDAQ 100.

C1825: **Detecting mean-field in a financial network model**
Presenter: **Tomoyuki Ichiba**, University of California Santa Barbara, United States
Starting with a finite system of linear stochastic equations for the nodes in a financial network, its limit is studied as the number of nodes goes to infinity. Then, we shall consider the system with Lipschitz continuous coefficients in the limiting system. The limiting system can be similar to the mean-field limit for which each node is attracted by the representative, mean node, but is now described by a system of equations where the

coefficients depend on another neighboring node which has the identical law. By comparing it with the mean-field limit, we discuss a detection problem of mean-field component in such financial systems.

C1613: **Inference for tangency portfolio weights for small sample and singular covariance matrix**
Presenter: **Joanna Tyrcha**, Stockholm University, Sweden
Co-authors: Taras Bodnar, Stepan Mazur, Krzysztof Podgorski
Estimation of covariance matrix between portfolio assets plays a very important role in risk management. The inverse of an estimate is used to estimate the optimal portfolio weights. The sample covariance matrix is typically used for this purpose under the assumption of a non-singular true (population) covariance matrix. However, the problem of potential multicollinearity and strong correlations of asset returns results in clear limitations in taking such an approach due to potential singularity or near singularity of the population covariance. Further, realistic financial applications are often characterized by a smaller number of observations (sample size) than the number of assets in the portfolio. A stochastic representation of the tangency portfolio weights estimator as well as the linear hypothesis test for the portfolio weights is presented under singular conditions caused both by the number of data points being lower than the number of assets and by the singularity of population covariance matrix. The asymptotic distribution of the estimated portfolio weights is also established under a high-dimensional asymptotic regime. The motivation for the considered singularities in real data and practical relevance of theoretical results are presented. The theoretical results are applied to real stock returns in an illustrative example.

CO539 Room MAL 415 CONSUMER CREDIT RISK Chair: Tony Bellotti

C0789: **Profit scoring in peer to peer lending**
Presenter: **Trevor Fitzpatrick**, University of Southampton, United Kingdom
The purpose is profit scoring of Peer to Peer loans. To accomplish this, a comparison of ensemble algorithms such as Regularised Random Forest and Regularised Regression is carried out. Transformations of the response variable are also investigated. The results suggest that rank-based transformations and ensemble-based regressions produce the most promising results.

C0881: **Clustering methods for consumer credit risk modelling**
Presenter: **Yazhe Li**, Imperial College London, United Kingdom
Co-authors: Niall Adams, Tony Bellotti
For credit risk modelling, we propose applying automated clustering on the class of default loans as a precursor to building a credit risk model using logistic regression. We illustrate an application of this method using simulated and real data sets, mortgages and credit cards, using k-means, k-medoids and hierarchical clustering. Our results show that clustering can enhance predictive performance when defaults can be demonstrably clustered into well-separated clusters. We explain the motivation by asymptotic results demonstrating that logistic regression only uses the rare class data points via the rare class mean vector in highly imbalanced data problems, such as default modelling.

C1361: **Social networks analytics using GOTCHA: Theory and applications**
Presenter: **Bart Baesens**, KU Leuven, Belgium
A common assumption in analytical applications is that customer behavior is independent and identically distributed, often referred to as the IDD assumption. However, in many real-life settings this assumption is simply not valid. Social network effects between customers, both implicit and explicit, create collective correlational behavior which needs to be appropriately analyzed and modeled. We will start by outlining the architecture of a social network learning environment, consisting of a local model (e.g. a logistic regression model), a relational learner (e.g. a relational neighbor classifier), and a collective inferencing procedure (e.g. Google PageRank). We will then introduce our recently developed GOTCHA method for social network analytics and will illustrate the application thereof in various real-life settings such as churn prediction, credit scoring and fraud detection. It will be empirically shown how GOTCHA allows us to efficiently model social network effects, hereby generating both additional lift and profit compared to, e.g., a flat logistic regression model.

C0867: **A binary spatial autoregressive sample selection approach for modeling access to finance for UK SMEs**
Presenter: **Michaela Kesina**, ETH Zurich, Switzerland
Co-authors: Raffaella Calabrese
Data in empirical applications often suffer from non-random draws from the population - an observed sample is rather the result of some selection mechanism. It is well-known that ignoring sample selection, where present, leads to biased and inconsistent parameter estimates. The spatial econometrics literature mostly focuses on spatial sample selection model variants, which allow for spatial correlation in the errors, to improve the efficiency of an estimator. However, ignoring spatial correlation in the dependent variable, where present, also leads to biased and inconsistent parameter estimates. Therefore we propose a binary spatial autoregressive sample selection model where we allow for spatial correlation in the dependent variable of both the selection and the outcome equation to model the interdependence in the respective choices. We estimate the model with a Bayesian Markov Chain Monte Carlo (MCMC) procedure and apply it to data on UK small and medium sized enterprises (SME) to investigate the access to finance.

C1832: **Incorporating heterogeneity and macroeconomic variables into multistate delinquency models for credit cards**
Presenter: **Jonathan Crook**, University of Edinburgh, United Kingdom
Co-authors: Viani Djeundje
Multistate delinquency models model the probability that a credit account transits from one state of delinquency to another between any two points in the life of the account. Using a large sample of credit card accounts we parameterise such models and investigate whether predictive accuracy is enhanced by the incorporation of account specific random effects, the use of B-splines for the baseline functions and the incorporation of macroeconomic variables. We conclude that macroeconomic variables are statistically significant in such models, that the inclusion of random effects renders the fixed effects less statistically significant but does not enhance predictive accuracy.

CO324 Room MAL 416 MACROECONOMETRICS Chair: Toshiaki Watanabe

C0272: **Reducing dimensions in a large TVP-SVAR**
Presenter: **Rodney Strachan**, The University of Queensland, Australia
Co-authors: Joshua Chan, Eric Eisenstat
A new approach is proposed to estimate high dimensional time varying parameter structural vector autoregressive models (TVP-SVARs) by taking advantage of an empirical feature of TVP-(S)VARs. TVP-(S)VAR models are rarely used with more than 4-5 variables. However recent work has shown the advantages of modelling VARs with large numbers of variables and interest has naturally increased in modelling large dimensional TVP-VARs. A feature that has not yet been utilized is that the covariance matrix for the state equation, when estimated freely, is often near singular. We propose a specification that uses this singularity to develop a factor-like structure to estimate a TVP-SVAR for 15 variables. Using a generalization of the recentering approach, a rank reduced state covariance matrix and judicious parameter expansions, we obtain efficient and simple computation of a high dimensional TVP-SVAR. An advantage of our approach is that we retain a formal inferential framework such that we can propose formal inference on impulse responses, variance decompositions and, important for our model, the rank of the state equation covariance matrix. We show clear empirical evidence in favour of our model and improvements in estimates of impulse responses.

C0474: **Identifying oil price shocks and their consequences: Role of expectations and financial factors in the crude oil market**
Presenter: **Jouchi Nakajima**, Bank for International Settlements, Switzerland
Co-authors: Takuji Fueki, Hiroka Higashi, Naoto Higashio, Shinsuke Ohyama, Yoichiro Tamanyu
A simple but comprehensive structural vector autoregression (SVAR) model is proposed to examine the underlying factors of oil price dynamics by explicitly incorporating the role of expectations on future aggregate demand and oil supply as well as financial investors role in the crude oil market. Our main findings are threefold. First, our empirical analysis shows that shocks on expectations and financial factors in the oil market explain more than 40 percent of historical oil price fluctuations. In particular, expected future oil supply shocks are more than twice as important as realized and expected aggregate demand shocks or financial factor shocks in accounting for the oil price developments. Second, focusing on a recent large drop in oil prices since 2014, the analysis reveals that expected future oil supply shocks were the dominant driver of oil price falls from January 2014 to January 2015, while expected and realized aggregate demand shocks played a major role in oil price falls from June 2015 to February 2016. Finally, we show that the influence of oil price shocks on global output varies by the nature of each shock.

C0899: **Cyclical part-time employment in an estimated new Keynesian model with search frictions**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Toshihiko Mukoyama, Kazuhiro Teramoto
The dynamics of full-time employment and part-time employment over the business cycle are analyzed. We first document the basic macroeconomic facts on these employment stocks using the U.S. data, and decompose their cyclical dynamics into the contributions of different flows into and out of these stocks. Second, we develop and estimate a New Keynesian search-and-matching model with two labor markets to uncover the fundamental driving forces of the cyclical dynamics of employment stocks.

C0693: **Fiscal confidence shocks and the market for the Japanese government bonds**
Presenter: **Etsuro Shioji**, Hitotsubashi, Japan
The question of fiscal sustainability is an important issue in many advanced countries, including Japan. With the mounting public debt, any news that could change general perceptions about government solvency could potentially have a large impact on the economy. We construct a new list of dates on which such news arrived in Japan, for the period between 2008 and 2017. This is done through careful readings of two major newspapers, Nikkei and Asahi. News can be classified into a few categories: (1) those related to introduction (or postponement) of a higher consumption tax rate; (2) those related to efforts to contain social security spending; (3) those about new big public construction plans; (4) governmental projections about long run fiscal balances; and (5) others. We then study what kind of effect (if any) those news have had on the market for the Japanese Government Bonds (JGBs), represented by daily returns on JGB futures and the JGB-VIX.

C0940: **Effects of corporate bond purchases and their transmission mechanism: The case of Japan**
Presenter: **Yoichi Ueno**, Bank of Japan, Japan
Co-authors: Kenji Suganuma
The purpose is to examine the effect and transmission mechanism of corporate and government bond purchases by the Bank of Japan on Japanese firms credit spreads. Exploiting both individual bond data and firm level data, we find that the bond purchases have significant effects on lowering the spreads through various channels: default risk premium channel, local and global supply channels, and risk taking channel. We also find that the risk appetite of financial intermediaries change the magnitude of the effects through the supply channels. Quantifying the effects on the spreads in various ratings or maturities, we find that the purchases of corporate and government bond have influence more on lower-grade and longer-maturity credit spreads. Furthermore, given the same amount of purchase, the corporate bond purchase has larger effect on lowering the spreads than the government bond purchase.

CO274 Room MAL 539 UNCERTAINTY IN MACRO-ECONOMIC NOWCASTING AND FORECASTING Chair: Gian Luigi Mazzi

C0222: **Inflation forecasts, inattentiveness and uncertainty**
Presenter: **Roberto Golinelli**, University of Bologna, Italy
Co-authors: Joshy Easaw, Saeed Heravi
The purpose is to investigate the nature of professionals inflation forecasts inattentiveness. We outline and empirically investigate a generalized model of inattentiveness due to informational rigidity. We also introduce a novel theoretical model that considers the relationship between inattentiveness and uncertainty and, crucially, distinguishing between macroeconomic and data uncertainty. The empirical analysis uses the Survey of Professionals Forecast (SPF) data and clearly indicate that professional forecasters are depicted by inattentiveness due to imperfect information.

C0530: **Improving accuracy of early quarterly estimates of GDP components in the U.S. national accounts**
Presenter: **Baoline Chen**, Bureau of Economic Analysis, United States
Differences in the source data used to compile the three quarterly GDP releases underlie a substantial part of the revisions between the first two releases and the third. Such revisions can be sizeable for some GDP components, especially during the periods of rapid changing macroeconomic conditions. Revisions to the early releases could also be due to a lack of sufficient relevant monthly information and a lack of information on the longer term dynamics of the GDP component series. The aim to develop a structured approach for compiling early estimates of GDP components in the national accounts, combining information on the longer-term characteristics of the component series and short-term movements in the series resulting from the changing economic conditions. The proposed estimation approach is primarily based on the bridge-equation and bridging with factor model in the now-casting literature, using a set of relevant current and lagged monthly indicators to provide more accurate information on the component series in the current quarter while employing lagged quarterly information to provide longer term dynamics for the component series.

We apply this approach to the detailed component series in the PCE of services with the objective of minimizing the root mean squared revision (RMSR) statistic.

C0871: **Nowcasting with large, international data sets: On sparse priors**
Presenter: **Philipp Hauber**, Kiel Institute for the World Economy, Germany
Co-authors: Christian Schumacher

Factor models can summarize the comovements of a large number of variables and have proven useful in nowcasting and short-term forecasting of GDP growth. The main aim is to assess the importance of international variables for nowcasting national developments, an issue, which curiously has received relatively little attention in the academic literature so far. However, given the large number of potentially relevant variables at both the national and international level, the question arises whether all this information is useful for nowcasting or not. As such, we also contribute to the continuing debate on variable selection and the optimal size of factor models for forecasting applications. Rather than choosing variables ad hoc, we employ sparse priors on the factor model's loadings which can help to identify those business cycle indicators that essentially determine the factors, whereas irrelevant variables are sorted out. In an empirical exercise, we evaluate nowcasts of GDP for the Euro area and the United States from models that use only the respective national data as well as the combined data sets comprising of roughly 150 variables.

C0986: **New euromind: A model averaging approach**
Presenter: **Gian Luigi Mazzi**, Independent Expert, Luxembourg
Co-authors: Alessandro Giovannelli, Tommaso Proietti

Gross domestic product (GDP) is the most comprehensive and authoritative measure of economic activity. The macroeconomic literature has focused on nowcasting and forecasting this measure at the monthly frequency, using related high frequency indicators. The issue of estimating monthly gross domestic product is addressed by using a large dimensional set of monthly indicators, by pooling the disaggregate estimates arising from simple and feasible bivariate models that consider one indicator at a time, in conjunction to GDP or a component of GDP. The weights used for the combination reflect the ability to nowcast the original quarterly GDP component. Our base model handles mixed frequency data and ragged-edge data structure with any pattern of missingness. We estimate monthly GDP according to both the direct and the indirect approach. In the latter case we estimate sixteen GDP components, by output and expenditure type, by pooling the monthly estimates arising from all possible bivariate models and we aggregate them using a methodology, the annual overlap method, that is used in the production of the national accounts, and that ensures the consistency in cross-sectional aggregation with the published total GDP at market prices. Our methodology allows to assess the contribution of the monthly indicators to the estimation of monthly GDP, thereby providing essential information on their relevance. This evaluation has led to several interesting discoveries.

C0211: **Predictability of Euro area revisions**
Presenter: **Katharina Glass**, Aurubis AG, Germany

The predictability of revisions to Euro- area major macroeconomic variables is investigated using real-time data from the European Central Bank. The application of nonparametric and semiparametric tests enables robust conclusions about the predictability of revisions. Though there is wide evidence of the nonnormality of the distribution function of revision errors, this is the first application of the nonparametric approach to examine revisions. Moreover, to gain robustness, tests for parameter instability are performed, and structural breaks are explicitly included in the predictability evaluation. The results underline the predictability of Euro area key macroeconomic revisions. Revisions are inefficient and biased, and revision errors are not optimal forecast errors.

CC711 Room MAL 152 CONTRIBUTIONS IN RISK ANALYSIS Chair: Enrico Biffis

C1508: **Parameter estimation and bias correction in the Vasicek credit portfolio model**
Presenter: **Marius Pfeuffer**, University of Erlangen-Nuremberg, Germany

Credit portfolio modeling is an important aspect of quantitative risk management and the Vasicek model with asset variables $A_{ij} = \sqrt{\rho_i}\Psi_i + \sqrt{1-\rho_i}\varepsilon_j$, where $\Psi \sim \mathcal{N}(\mathbf{0}, \{\rho_{i_1,i_2}\}_{1\leq i_1,i_2\leq I})$ is a systematic and $\varepsilon \sim \mathcal{N}(0,1)$ an idiosyncratic component is a core approach to describe dependencies for a set of obligors $1,\ldots,J$ in cohorts (e.g., rating categories) $1,\ldots,I$. In order to parameterize this model, moment or maximum likelihood estimators for the correlation parameters are usually employed. First, a systematic overview of the existing estimation approaches in the literature is given. Second, as empirical studies report that asset correlation estimates are often biased, this effect is analyzed and possible methods for bias reduction are discussed. It is especially shown that what in the literature is considered as moment estimators is based on the asymptotic behavior of the Binomial-Vasicek compound distribution. The moment estimators for the complete model are derived and its behavior is illustrated. Moreover, resampling approaches for bias reduction are employed. Third, theoretical properties of inter-asset Pearson correlations ρ_{i_1,i_2} between different cohorts are outlined and benchmarked against other measures of dependence. The methods are illustrated by hypothetical simulation studies and Standard and Poor's ratings data.

C1550: **Quantifying effects of extreme events with applications to financial crises**
Presenter: **Alexander Ristig**, University of Vienna, Austria
Co-authors: Ostap Okhrin, Jeffrey Sheen, Stefan Trueck

The aim is to address the quantification of effects from an extreme outcome in an explanatory variable on a dependent variable. The effect is approximated with the so-called asymptotic elasticity of a conditional quantile function, linking the dependent and explanatory variable. A closed form expression for this asymptotic elasticity is presented under reasonable assumptions on the exact relation between explanatory and dependent variable. By interpreting the asymptotic elasticity as a spill-over measure for tail-risk, we detect statistically significant effects from Lehman Brothers to other financial institutions during the subprime mortgage crisis before Lehman Brothers was obviously in distress. Likewise, the effect from a credit default in case of Greece on the solvency of countries within the Euro-area is briefly studied.

C0180: **The tail behaviour due to the risk premium in AR-GARCH-in-mean, GARCH-AR and double-autoregressive-in-mean models**
Presenter: **Emma Iglesias**, University of A Coruna (SPAIN), Spain
Co-authors: Christian Dahl

Results in extreme value theory are extended by describing the tail behaviour when a risk premium component is added in the mean equation of different conditional heteroskedastic processes. We study three types of parametric models that are able to generate a risk premium: the traditional GARCH-M, the double autoregressive model with risk premium and the GARCH-AR model. We find that if an autoregressive process is introduced in the mean equation of a traditional GARCH-M process, the tail behavior is the same as if it is not introduced. However, if we add an autoregressive process to a conditional volatility model with a risk premium component and lags of data in the conditional variance, then the tail behaviour changes. The GARCH-AR model also has a different tail index than the traditional AR-GARCH model. In our simulations, we show that the larger tail indexes are generated when using the traditional GARCH-M model. Also, when the risk premium increases, the tail index tends to fall with the only exception of specifying a risk premium with the logarithm of the volatility in the double autoregressive model. We also show some parameter settings where the strong stationarity condition of the risk premium models fails.

C1559: **Portfolio optimization based on dynamic factor and dynamic conditional correlation GARCH models**
Presenter: **Maziar Sahamkhadam**, Linnaeus University, Sweden
Co-authors: Andreas Stephan, Ralf Ostermark

Dynamic Factor Analysis (DFA) is used in the context of portfolio optimization based on Dynamic Conditional Correlation (DCC)-GARCH forecasting models. We use Gaussian State-Space modeling and Kalman filtering to extract the hidden state variables (factors), assuming the standard identifiable model and parameter constraints. The hidden factors are then inserted into the mean equation of DCC-GARCH and ARMA-DCC-GARCH models to perform out-of-sample forecasts. Having obtained mean and volatility forecasts, we simulate one-day ahead returns from multivariate normal distribution and allocate optimal weights based on three utility functions including maximum Sharpe ratio (CET), Global Minimum Variance (GMV) and minimum Conditional Value-at-Risk (Min-CVaR). We apply the DFA-DCC-GARCH and DFA-ARMA-DCC-GARCH models to portfolios which consist of twelve U.S. industry indexes or ten stock indexes and compare them with simple DCC-GARCH models as the benchmarks. In general, for both stock indexes and industry indexes, the models based on DFA outperform the benchmarks either in terms of maximizing the investor's utility function (maximizing Sharpe ratio or minimizing CVaR) or increasing the accumulation wealth. However, there is not much improvement for GMV portfolios. In conclusion, the DFA-based models are suitable for both CET and Min-CVaR portfolios.

C0204: **The impact of news on firm-specific risk-neutral higher moments**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States
Co-authors: Sirio Aramonte

A methodology is proposed to study the impact of arrival of news on risk-neutral higher moments of a firm's expected stock returns, by blending option prices and CDS spreads - with the former providing information about the central part of the distribution and the latter determining the left tail. We apply the methodology to a large sample of risky U.S. and European firms. We assess the economic impact of arrival of macroeconomic, corporate, and political news and their propagation through the second and third risk-neutral (option implied) moments. Our methodology is closely related to a recent work which combines CDS and options data to extract firm-specific risk-neutral distributions. It documents the economic value by carrying out time-series and cross-sectional asset pricing tests. Our study differs from that work in that a) our focus is on transmission of news rather than gauging the economic value of blending options and CDS data and b) we use international data in addition to U.S. data.

CC709 Room MAL 402 CONTRIBUTIONS IN FORECASTING Chair: Frederique Bec

C0388: **Non-linear causality test based on partial distance correlation: Application to energy futures**
Presenter: **German Creamer**, Stevens Institute of Technology, United States
Co-authors: Chihoon Lee

A nonlinear causality test is proposed by using the partial distance correlation. As an extension of the Brownian distance correlation, partial distance correlation ($R(X_{t-l}, Y_t; Y_{t-1})$) calculates the distance correlation between random variables X and Y given a Z random variable. Our test evaluates the non-linear causality of any time series such as the current value of Y (Y_t) on the l lagged values of X ($X_{t-1}...X_{t-l}$) given the past values of Y ($Y_{t-1}...Y_{t-l}$) using a modified version of the partial distance correlation $R(X_{t-1}...X_{t-l}, Y_t; Y_{t-1}...Y_{t-l})$. Our test uses moving block bootstrap (MBB) to compute the empirical p-values and selects the most relevant variables. The main reason to use MBB is to evaluate the effect of time dependence using variables that are independent and identically distributed.Our method determines relationships similar to those identified by the linear Granger causality test, and it also uncovers additional non-linear lagged relationships among the log prices of oil, coal, and natural gas futures. When these linear and non-linear relationships are used to forecast energy futures with a non-linear regression method such as support vector machine for regression, the forecast of energy futures improves when compared to a forecast based only on Granger causality and by a baseline autoregressive model selected using the Akaike information criterion.

C1471: **Forecasting fiscal aggregates in an emerging market economy: The role of macro variables and fiscal plans**
Presenter: **Antonio Jesus Sanchez Fuentes**, University Complutense of Madrid, ICEI and GEN, Spain
Co-authors: Javier J Perez, Engin Yilmaz

Against the framework of recurrent fiscal vulnerabilities in emerging market economies, a modeling approach is proposed to forecast and monitor fiscal policy developments in the short-run for the case of Turkey. We pose a suite of models which combine fiscal and macroeconomic indicators with annual fiscal policy plans, by using mixed-frequencies, time series models estimated at the monthly frequency. This type of approach has been used successfully for the case of developed economies. We carry out a recursive pseudo real-time estimation to compare the performance of these models, with the aim of comparing their accuracy and determining the role of macroeconomic variables, as well as the information content of (ex ante) fiscal plans.

C1797: **Modelling and forecasting Euro area GDP growth using a hierarchical factor model based 3PRF**
Presenter: **Rolf Scheufele**, Swiss National Bank, Switzerland
Co-authors: Alain Galli, Massimiliano Marcellino

A multi-level factor model is set up for analyzing and forecasting GDP growth for the aggregate Euro area and its largest members (Germany, France, Italy and Spain) in one aggregate framework. This framework is used to characterize fluctuations in real activity of the aggregate and the individual countries. It allows us to decompose each country's GDP growth into a global (Euro area wide) component, a country specific-component and an idiosyncratic component. To extract the factors of the different levels, we employ a principle component based quasi-maximum likelihood using targeted predictors as well as a three-pass regression filter (3PRF) approach. The results suggest that the three-pass regression filter approach is able to explain a larger country-specific share of GDP growth than the standard quasi-maximum likelihood approach. Moreover, it turns out that our proposed 3PRF based hierarchical factor model performs similarly than an aggregate factor model for the Euro area before the outbreak of the European debt crisis, but notably better thereafter.

C1565: **Forecasting US inflation using Markov dimension switching**
Presenter: **Jan Prueser**, Ruhr Graduate School in Economics, Germany

Bayesian variable selection in the Phillips curve context is considered by using the Bernoulli approach. The Bernoulli model, however, is unable to account for model change over time, which is important if the set of relevant predictors changes over time. To tackle this problem, the Bernoulli model is extended by introducing a novel modeling approach called Markov Dimension Switching (MDS). MDS allows the set of predictors to change over time. The MDS and Bernoulli model reveal that the unemployment rate, the Treasury bill rate and the number of newly built houses are the most important variables in the generalized Phillips curve. Furthermore, these three predictors exhibit a sizeable degree of time variation for which the Bernoulli approach is not able to account, stressing the importance and benefit of the MDS approach. In a forecasting exercise the MDS model compares favorably to the Bernoulli model for one quarter and one year ahead inflation. In addition, it turns out that the forecasting performance of MDS model is competitive in comparison with other models found to be useful in the inflation forecasting literature.

C1561: **Exploit market microstructure noise in volatility forecasting**
Presenter: **Ye Zeng**, CREATES, School of Business and Social Science, Aarhus University, Denmark, Denmark

In volatility forecasting, realized volatility as an estimator for the latent quadratic variation of asset price unavoidably begets the measurement error in variable problem. In a similar vein to the HARQ model which tackles the problem by attenuating realized volatility according to its asymptotic variance, another simple extension of the heterogenous autoregressive (HAR) model, named HARN, is proposed which exploits size of market microstructure noise to gauge reliability of realized volatility. Empirical analysis on datasets covering 29 stocks listed in NYSE show that realized volatility is always attenuated in response to the noisiness of data. Improved forecasting accuracy is also documented, both in-sample and out-of-sample, in comparison with results of the standard HAR model. In addition, empirical results show that the HARN model utilizing a simple estimator for the size of noise computed with data sampled every 5 minutes outperforms, or at least is on par with, models incorporating more sophisticated estimators. Thus, by augmenting HAR model with an extra simple noisiness measure, we obtain a parsimonious extension which improves volatility forecasting.

CC712 Room MAL 421 CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: Thomas Beissinger

C1413: **Estimating a time-varying financial conditions index for South Africa**
Presenter: **Alain Kabundi**, South African Reserve Bank, South Africa

A financial condition index for South Africa is estimated by using 41 monthly time series of financial market observed from January 2000 to April 2017. The empirical technique used is a dynamic factor model with time-varying factor loadings based on principal component analysis and the Kalman smoother. In addition, we estimate a time-varying factor-augmented vector autoregressive (TVP FAVAR) which includes in addition to the FCI two observed macroeconomic variables. The results show the ability of estimated FCI to predict risks in the financial market emanating from the domestic market and the global market. Furthermore, TVP FAVAR outperforms the constant-loadings FAVAR and the traditional VAR in out-of-sample forecasting of inflation rate and the real GDP growth rate. Finally, the shocks from the FCI affect the real GDP growth rate negatively, whereas the inflation rate reacts positively. Importantly, the responses of macroeconomic variables vary over time.

C1835: **Reassessing competitiveness at the sectoral level: A new unit labor cost indicator based on value-added chains**
Presenter: **Martyna Marczak**, University of Hohenheim, Germany
Co-authors: Thomas Beissinger

A new international competitiveness measure at the sectoral level is proposed that is based on unit labor costs and takes into account value added chains. Our cost-based indicator is more suitable than often employed export price-based indicators if firms employ a pricing-to-market strategy. Since in modern economies with strong intersectoral linkages a sectors total costs to a large extent depend on costs of inputs from other industries, we incorporate these linkages to obtain the so-called embedded labor costs. More specifically, they are given by a weighted average of unit labor costs of all domestic sectors contributing to the output of a specific sector, where the contribution of each domestic sector is derived from international input-output tables. We justify this sectoral competitiveness measure with a simple Leontief-type model. As an example, we calculate this measure for German sectors and show that, compared to standard competitiveness measures, it more clearly reflects the often stated argument that German manufacturing industries experienced a significant increase in competitiveness since the 90's.

C1503: **An analysis of ASEANs logistics performance and its impact on regional trade: An extended gravity model approach**
Presenter: **Barton Sy**, TAT SING International Logistics Corporation, Philippines
Co-authors: Stephen Jun Villejo

The aim is to investigate the impact of logistics, using the Logistics Performance Index, on bilateral trade among ASEAN nations using the gravity model on a panel data on certain years from 2007 to 2016. The overall LPI index, and its components, are strong contributors to a healthy trading among countries in the ASEAN region. The ranking of the importance of the components of the logistics index differ between the importing and exporting countries in general. It can also be observed that a dynamic model rather than a static model is more appropriate since the parameter estimates are observed to be changing through time. Instrument Variable Generalized Method of Moments is used based on the result of the Hausman test and in order for time-invariant variables to be retained in the model. Since of particular interest of the researcher is the Philippines, the result of the model is used to investigate which countries in the region the Philippines has the highest trade potential with. Due to an established relationship among ASEAN countries, there is an insignificant trade potential for all countries, using the method of speed of convergence.

C1781: **Revisiting the impact of trade openness on informal and irregular employment in Egypt**
Presenter: **Melika Ben Salem**, University Paris-Est Marne-la Vallee, France
Co-authors: Chahir Zaki

The impact of trade openness on job quality is examined through the evolution of the shares of informal and irregular employment in total employment. In fact, Egypt has undertaken several liberalization waves and reforms of the labor market (1998-2012). Moreover, the economy has been subject to several events leading to a severe political instability which in turn affected production, exports, employment and employment conditions. Indeed, informal and irregular employments have exacerbated in the wake of the political turmoil of 2011. Thus, combining a microeconomic dataset (the Egyptian Labor Market Panel Survey) with macroeconomic variables (tariffs), we try to assess to what extent trade reforms affected informal/irregular workers in Egypt. Our main findings show that there is a positive association between tariffs and both informal and irregular employments in Egypt. While the effect on informality is robust, the one on irregularity is not.

E1721: **Computation of three discrete mixtures of continuous distributions: Stability analysis**
Presenter: **Joana Leite**, ISCAC-IPC and CMUC, Portugal
Co-authors: Jose Carlos Dias, Joao Pedro Nunes

The distributions of the squared sample multiple correlation coefficient, noncentral t, and noncentral chi-square are critical in many fields, including Finance. For example, the latter has applications in the context of interest rate modelling, valuation of financial and real options, and the simulation of stochastic volatility processes. Previous algorithms for computing these distributions achieved significant gains in efficiency when compared to the previous ones, and, nowadays, are considered the benchmark. In their core are recurrence relations of special functions, applied in both the forward and in the backward directions. Recently, stability issues have been raised regarding the algorithm for the noncentral chi-square distribution. In this setting, all three algorithms are analysed, with particular attention paid to the stability of the recurrence relations. Necessary modifications are implemented in the algorithms and comparison studies carried out.

CC706 Room MAL 538 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II Chair: Arvid Raknerud

C0356: **Multidimensional time-frequency analysis of the CAPM**
Presenter: **Roman Mestre**, University of Montpellier, France
Co-authors: Michel Terraza
The CAPM theory provides a measure of the sensitivity of an asset to the market called the systematic risk. The Beta is estimated by its market line. According to the OLS hypothesis, it is stable over time but this is not empirically verified. Many studies support this fact, and more particularly the Beta dispersion according to the frequencies which is related to the heterogeneous behavior of agents. Using the wavelets method, we can calculate the coherence and the phase between the stock's returns and those of the market over time. In order to confirm the correctness of the methodology, we use three French equities with different Betas (AXA, LVMH and Orange) from 2005 to 2015 period including the crisis. We show that the wavelets coherence, associated with the phase, improves our understanding about the Equity-Market relationship and then, the classification of equities according to their frequency characteristics (the contagion and interdependence phenomenon). The contagion effects (from the market to the stock) is principally located on the High Frequencies whereas the interdependence effect is located on the Low-Frequencies (Long-run investment). The link between beta and coherence-phase can help the investors to choose more efficiently the time they should invest.

C0582: **The asymmetry in carry trade and the U.S. dollar**
Presenter: **Chang-Che Wu**, Yuan Ze University, Taiwan
Co-authors: Chih-Chiang Wu
The aim is to investigate the potential asymmetric dependence between the carry trade and U.S. dollar returns. Empirical results demonstrate that the U.S. dollar becomes a safe haven and provides protection for carry trade investors to avoid the crash risk during the 2007-2008 global financial crisis and the 2010-2011 Eurozone sovereign debt crisis. The asymmetric dependence is not only statistically significant, but this information also helps investors to generate extra 14-2,166 annualized basis points from the perspective of an asset-allocation decision. Our findings provide important financial implications for currency investors in asset allocation and risk management.

C1527: **Monthly forecasting of the dollar to the ruble exchange rate. Adaptive Kalman filter**
Presenter: **Sergei Borodachev**, Ural Federal University, Russia
The goal is to develop a model that allows us to forecast the dollar to the ruble exchange rate for a month ahead based on macroeconomic data, published at monthly intervals. Proposed structural model of the dynamics of the ruble and dollar masses that determine the exchange rate, depending on changes in foreign exchange reserves, the balance of foreign trade, the monetary base, the MICEX index, the price of oil. With the help of the Kalman filter, the model parameters, the dynamics of the money masses, and the forecasting of the dollar exchange rate were estimated. Monthly data were used from the beginning of 2015 to mid-2017. The estimation of the capacity of dollar market was found in about half the capacity of the MICEX index funds.Average error of forecasts, based on information available one step before the forecasted moments (RMSEA) was 1.99 RUB. Adaptive form of KF was developed when, similarly to the EM algorithm, the phases of KF estimation in the window and minimization of average prediction error to determine the optimal estimates for the system model parameters in this moment are sequentially alternated. With this RMSEA became 1.39 RUB.

C1718: **Modeling interest rate pass through with heterogeneous bank loans**
Presenter: **Arvid Raknerud**, Statistics Norway, Norway
Co-authors: Bjorn Helge Vatne, Paul Mizen
Studies of interest rate pass-through make a common assumption that bank loans are homogenous products priced in an imperfectly competitive framework. We explore the implications of relaxing that assumption. We argue that, in principle, the products banks offer their customers are heterogeneous and not so different from the goods and services offered by the non-bank private sector; therefor rate setting could be modeled using a modified Dixit-Stiglitz model. Funding costs are determined by money market rates and spreads on bank bonds that vary across banks, whereas retail rates are mark-ups over marginal funding costs. Applying our model to data from Norway, our results regarding long-term pass-through show that it is incomplete, with a 100bp increase in the market rate leading to a 89bp increase in household mortgage rates and a 79bp increase in business loan rates on average. We also find that forward NIBOR rates and spreads on bonds have a significant impact on retail rate setting, with corporate loan rates varying much more across banks than mortgage rates do. Almost 60 percent of the between-bank variation in corporate loan rates and 40 percent in household mortgage rates can be explained by bank-fixed effects, whereas, for both types of loans, almost 20 percent of the between-bank variation in lending rates is due to either observed differences in spreads or differences in pass-through coefficients across banks.

C1848: **On lead-lag correlations in stochastic volatility models with jump**
Presenter: **Sujay Mukhoti**, Indian Institute of Management Indore, India
Co-authors: Pritam Ranjan
In an efficient market, the returns and their time dependent volatility are often jointly modelled by stochastic volatility models (SVMs). SVM posits the current return as a function of current latent volatility in a hierarchical model. Volatility is further modelled as an auto-regressive process. In an attempt to include the opposite reaction of return to its volatility, the return and volatility shocks are assumed to be negatively correlated (SVL-I), which however leads to violation of the efficient market hypothesis. Subsequently, a host of alternative SVMs (SVL-II) had been developed where current return shock is assumed to be correlated with future volatility shock. On the contrary, recent empirical works provides support for a strong contemporaneous correlation between return and volatility. The SVL-II models, however, fails to account for such contemporaneous correlation. We propose a correction of the SVL-I type models to satisfy EMH and capture the contemporaneous correlation as well. We also establish theoretical results in support of empirical lead-lag correlation pattern between return and volatility. We inspect similar results in case of SVM with leverage and jumps. Furthermore, two real-life examples (Euro-USD rate, and S&P 500 index returns) have been used to demonstrate the performance of this new class of SVMs.

CC714 Room MAL 540 CONTRIBUTIONS IN FINANCIAL APPLICATIONS Chair: William Dunsmuir

C1443: **Evaluation and analysis of the value of German real estate following the financial crisis of 2007**
Presenter: **Chong Dae Kim**, Technical University of Cologne and Witten/Herdecke University, Germany
The price of housing is an important indicator in the analysis of a macroeconomy. In the last fifty years the development of the housing market in all large industrial countries showed a collapse before each large recession. The calculation of a property's value occurs through a partial solution of the regression equation and the use of the norming principle of real estate evaluation. Thus using this method we calculate the relative value of real estate for the German market and analyze the effect of the contemporaneous interest rate on the real estate market during the last ten years. Our results show that a low interest rate has had a positive effect on real estate prices in large cities as well as in district cities. Our results also show that prices in rural areas have been rising since 2013.

C1564: **Properties of the Margrabe best-of-two strategy to tactical asset allocation**
Presenter: **Giang Nguyen**, Vrije Universiteit Brussel, Belgium
Co-authors: David Ardia, Kris Boudt, Stefan Hartmann
The Margrabe Best-of-two strategy is a rule-based dynamic investment solution for the two-asset allocation problem. Its typical implementation involves yearly rebalancing the portfolio weights to 50-50 between a high-risk and low-risk asset. It uses intra-year weight adjustments to chase the momentum of the best performing asset by replicating the value of a Margrabe option to exchange an asset for another. In practice, this means that the Margrabe portfolio allocation benefits from the upside potential of the high-risk asset and the downside protection from the low-risk asset. The strategy depends on the assets prices, their return volatilities and correlation as well as the remaining time until year-end. We derive analytical formulae and use simulations to provide insights on the sensitivity of the strategy's weights and performance to these input parameters. We also report the results of an extensive out-of-sample evaluation for the bond-equity investment problem, where we compare the strategy with alternative solutions.

C1747: **Anatomy of the Eurozone crisis: A survival approach**
Presenter: **Haluk Yener**, Istanbul Bilgi University, Turkey
Two problems related to growth and survival probability maximization of an economy are solved. We assume that the problems are subject to a stochastic net worth model and apply the techniques of stochastic optimal control theory in order to find our results. From the results, we first construct a normative model with a dynamic threshold that designates the borrowing potential of an economy and a parameter that is used to see when borrowing more than the potential poses threat to an economy. In essence, we build a model that spots the non-resilient periods of an economy and introduce an indicator that demonstrates the level of non-resilience given the heterogeneous effect of crisis in Eurozone. Secondly, we aim to understand why economies exceed their borrowing potential by building an analytic relationship between the leverage strategy of an economy and its economic condition. Given this, we then construct a vector autoregressive (VAR-) model to show why certain economies borrow more than they should and eventually become non-resilient.

C1688: **Dynamic correlation network analysis of financial asset returns with network clustering**
Presenter: **Takashi Isogai**, Tokyo Metropolitan University, Japan
A novel approach is proposed to analyze a dynamic correlation network of highly volatile financial asset returns by using a network clustering algorithm to deal with high dimensionality issues. We analyze the dynamic correlation network of selected Japanese stock returns as an empirical study of the correlation dynamics at the market level. Two types of network clustering algorithms are employed for the dimensionality reduction. Firstly, several stock groups instead of the existing business sector classification are generated by the hierarchical recursive network clustering of filtered stock returns to overcome the high dimensionality problem due to the large number of stocks. Those group-based stock returns are filtered in advance to control for volatility fluctuations that can distort the correlation between stocks. Thus, the correlation network of individual stock returns is transformed into a correlation network of group-based portfolio returns. Secondly, the reduced size of the correlation network is extended to a dynamic one by using a model-based correlation estimation method. A time series of adjacency matrices is created on a daily basis as a dynamic correlation network from the estimation results. Then, the correlation network is summarized into only three representative correlation networks by clustering along the time axis. Some intertemporal comparisons are conducted by examining the differences between the three sub-period networks.

C0279: **Hypergeometric functionals and kernel regression in risk neutral density estimation**
Presenter: **Ana Monteiro**, University of Coimbra, Portugal
Co-authors: Antonio Santos
The analysis of financial risk through the information contained in options prices has been for long time considered important, and one way to extract such information is through the indirect estimation of the risk-neutral densities (RND). We compare different methods to retrieve information from estimated RND. There are several approaches for the estimation: parametric, semi-nonparametric and nonparametric methods, in particular, we consider the use of hypergeometric density functionals, and less structured methods like the ones based on kernel regression density estimation. We address the computational challenges associated with the use of hypergeometric functions, and the need of different amounts of data. The comparisons are made through simulated data and also from intraday market data, which to our knowledge is a novel approach. Due to the fact that the RND is not observable, a simulation analysis will attest that the methods used are able to capture the "true" density (in case of simulated data), and afterwards both methods are applied to real data sets. The root mean integrated error (RMISE) criterion is used to assess the quality of the estimation.

CG113 Room MAL 151 CONTRIBUTIONS IN PORTFOLIO OPTIMIZATION Chair: Nikolas Topaloglou

C1568: **Optimal asset allocation with heterogeneous persistence of shocks**
Presenter: **Federico Severino**, Universita della Svizzera Italiana, Switzerland
Co-authors: Fulvio Ortu, Claudio Tebaldi, Domenica Di Virgilio
There is wide evidence that financial time series are the outcome of the superposition of processes with heterogeneous frequencies. This is true, in particular, for market return. Indeed, log market return can be decomposed into uncorrelated components that explain the reaction to shocks with different persistence. The instrument that allows us to do so is the Extended Wold Decomposition, developed in a previous work. Hence, we construct portfolios of these components in order to maximize the utility of an agent with a fixed investment horizon. In particular, we build upon Campbell-Viceira solution of the optimal consumption-investment problem with Epstein-Zin utility, by using a rebalancing interval of 2^J periods. It comes out that the optimal asset allocation involves all the persistent components of market log return up to scale J. Such components play a fundamental role in characterizing both the myopic and the intertemporal hedging demand. Moreover, the optimal policy prescribes an increasing allocation on more persistent securities when the investor's relative risk aversion rises. Finally, portfolio reallocation every 2^J periods is consistent with rational inattention. Indeed, observing assets value is costly and transaction costs make occasional rebalancing optimal.

C0520: **Stochastic spanning and investment opportunities**
Presenter: **Sofia Anyfantaki**, Athens University of Economics and Business and Bank of Greece, Greece
Co-authors: Nikolas Topaloglou, Stelios Arvanitis
The notion of stochastic spanning is generalized to avoid uniformity over a class of investor preferences. This could be useful for developing tests for stochastic dominance efficiency with better properties. We derive an analytical characterization for stochastic spanning and the null limit distribution for the associated empirical test statistic. We also develop a subsampling procedure for the approximation of the asymptotic critical values and show asymptotic exactness when the significance level is restricted to a continuity point of the limiting quantile function and the null limit distribution is not degenerate. Consistency is also derived. An approximation that is numerically implementable is also provided. We employ the stochastic spanning approach to the question of whether to include cryptocurrencies in investors' portfolios. We further explain our results by documenting that cryptocurrency markets are segmented from the equity and bond markets.

C1716: **Aggregating strategies for online portfolio optimization**
Presenter: **Christine Keribin**, INRIA - Universite Paris-Sud, France

Online portfolio selection sequentially selects a portfolio over a set of assets, aiming to maximize the cumulative wealth. There is a large number of selection algorithms, that can be classified according to their leading principle: buy and hold, follow-the-winner, follow-the-user or pattern-matching approaches for example. The performance of these algorithms can depend on the market trend, some being more adapted in declining or in upraising period. These strategies can be viewed as experts, and they can be aggregated with online computed weights to propose a predictor than can be guessed better than the best expert. In this communication, we first compare standard strategies with a strategy based on a machine learning prediction. Then, we introduce aggregating algorithms and discuss their calibration. We compare the performance of the aggregation with these of the experts, and with prescient strategies. We examine the number of aggregated experts and the influence of the market trend. All the experiments are based on the CAC40 asset prices.

C1545: **Portfolio optimization based on forecasts from vine copula GARCH models using external regressors**
Presenter: **Andreas Stephan**, Jonkoping University, International Business School, Sweden
Co-authors: Maziar Sahamkhadam, Andreas Stephan, Ralf Ostermark

Based on AR-GARCH-Vine copula models, we forecast and simulate one-day-ahead returns of twelve industry portfolios. We add Fama-French factors to the mean and the volatility index (VIX) to the volatility equation as external regressors to the AR-GARCH model. For modeling the dependency structure between assets, we use regular, drawable and canonical vine copulas. Copula families include Clayton, Student-t and a mix of Gaussian, Student-t, Clayton and Gumbel with all rotations. We also consider skewed t and generalized extreme value distributions for modeling univariate marginals. We extend previous studies and provide estimates using the selection and estimation algorithm for regular vine copula. In general, the portfolio strategies based on maximizing the Sharpe ratio (CET) are outperforming the ones with minimizing conditional value-at-risk (MinCVaR). For CET portfolio strategies, skewed t distribution (as the marginal distribution) shows higher out-of-sample economic performance in comparison to generalized extreme value distribution (GEV). In terms of accumulated returns, portfolio strategies based on regular vine copulas are outperforming the strategies based on canonical and drawable vine copulas. In general, for MinCVaR portfolios, there is not much difference when changing vine copula family.

C1497: **Hybrid graphical least square estimation and its application in portfolio selection**
Presenter: **Saeed Aldahmani**, United Arab Emirates University, United Arab Emirates
Co-authors: Hongsheng Dai, Qiaozhen Zhang

A novel regression method based on the idea of graphical models is proposed to deal with the portfolio optimisation problem within the Markowitz mean-variance framework, when the number of assets V is larger than the sample size N. Unlike the regularisation methods such as ridge regression, LASSO and LARS, which give biased estimates,the newly proposed method can yield unbiased estimates for important variables, which contributes to improving the portfolio's Sharpe ratio by increasing its expected returns and decreasing its risk. Another characteristic of the new approach is that it produces a non-sparse portfolio that is more diversified in terms of stocks and reduces the stock-specific risk. To assess the proposed method, analysis of real data from London Stock Exchange and three simulation scenarios were carried out and the results were compared with those obtained from ridge regression. It is revealed that the proposed method significantly outperforms ridge regression in constructing portfolios with a higher Sharpe ratio for in-and-out-of-sample performance.

CG482 Room MAL 153 CONTRIBUTIONS IN MONETARY POLICY Chair: Matteo Ciccarelli

C1467: **Financial crises and optimal unconventional policies in international business cycles**
Presenter: **Shifu Jiang**, University of Surrey, United Kingdom

The Ramsey problem is studied for three unconventional policies in a two-country model with occasionally binding financial constraints. Lending to banks is the most efficient policy because of a smaller crowding out effect on private funds. If financed by government debt or a lump-sum tax, the optimal policy begins with a strong response and exits slowly. If financed by a labour tax, the unconventional policy is only used passively. A lack of international policy cooperation does not necessarily reduce welfare. A simple feedback rule that responds to gaps in asset prices mimics the optimal policy very well.

C0242: **To create or redistribute: That is the question**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus
Co-authors: Christos Savva, Nektarios Michail, Niki Papadopoulou

The aim is to explain the post-crisis low corporate investment is attempted to be explained despite aggressive easing in the financial conditions. Agents can utilize funding by either investing in capital or by redistributing existing assets. The former increases total income and employment while the latter alters the distribution of wealth among agents. We document empirically and explain theoretically that wealth redistribution increases while capital investment decreases during recessions because it is less risky for banks and more profitable for investors to redistribute existing wealth rather than create new. This exacerbates recessions and slows recoveries as it starves entrepreneurs from funding affecting capital creation. In addition, the aim is to explain rising inequality in recessions and why inequality can be harmful. As asset redistribution is a privilege of the rich, an increase in inequality encourages more income redistribution making recessions more severe. Macroprudential policies promoting access to finance to the "good-production" sector and discouraging asset redistribution can potentially boost recoveries.

E1692: **Revisiting the monetary policy asset price nexus by means of a novel statistical identification approach**
Presenter: **Hannes Rohloff**, University Goettingen, Germany
Co-authors: Helmut Herwartz

The aim is to revisit the monetary policy asset price nexus linked to the so-called lean against the wind debate that has recently gained attention. Typically, this issue has been investigated by means of structural VARs relying on a-priori assumptions which are often theoretically not fully supported or which are at odds with the data. We follow a different route and approach identification by detecting least dependent structural shocks under the presumption of non-Gaussianity. The statistically identified shocks coincide with meaningful economic counterparts, and allow a comparison with those of more conventional identification approaches. We provide empirical evidence on the US economy for monetary policy shocks and shocks originating from two asset markets: Equity and housing. The results indicate that contractionary monetary policy shocks have a mildly negative impact on both asset prices. The effect is less pronounced for equity. Vice versa, asset price shocks invoke a pronounced systematic policy response. Moreover, we find considerable differences in the speed of transmission of monetary policy shocks among both asset classes. On the contrary, the macroeconomic effects of asset price shocks appear similar.

C1547: **International monetary policy transmission**
Presenter: **Annika Schnuecker**, DIW Berlin, Germany
Co-authors: Gregor von Schweinitz

The purpose is to assess the macroeconomic effects of international monetary policy transmission for the United States, the United Kingdom and the Euro area. We use a Bayesian Proxy Panel SVAR in order to capture international linkages and to trace the dynamic responses of the

macroeconomic variables. A specific selection prior incorporating the panel dimension allows the estimation of the large number of parameters in the PVAR model. The monetary policy shocks of the three regions are identified via changes in daily future contracts around policy announcement dates. We use changes in stock prices as second proxies combined with sign restrictions to disentangle central bank information shocks from the monetary policy surprises.

C0510: **Re-assessing monetary policy shocks in China**
Presenter: **Bjoern van Roye**, European Central Bank, Germany
Co-authors: Alistair Dieppe, Paolo Bonomolo
The effects of monetary policy shocks on economic activity in China is investigated by using a variety of Bayesian VAR techniques. We show that monetary policy shocks appear to have a significant effect on economic activity under standard BVAR specifications. These findings are robust across different model specifications and different interest rate measures. However, this result changes when we consider a specification which takes into account the non-stationarity of the data. In particular, we consider an underlying economic model for the long-run with either a linear trend for GDP or with a common time-varying equilibrium on the steady-state jointly with foreign demand. In the latter case we find that the common structural explanation for the economic downturn crowds out the role of interest rates.

CG070 Room MAL 532 CONTRIBUTIONS IN COPULAS Chair: Richard Gerlach

C1421: **Testing for conditional dependence between domestic indexes using nonparametric copulas**
Presenter: **Jone Ascorbebeitia**, University of the Basque Country, Spain
Co-authors: Eva Ferreira, Susan Orbe
It is well known that the comovements between portfolios are time-varying. The interest is to detect in which level the comovements variation between domestic indexes can be explained by some global risk factors. Concretely, we will study the dependence of Euro Stoxx. Moreover, we are interested in measures of dependence beyond linear Pearsons correlation. This coefficient suits for normal variables, but not for variables with more complicated distributions such as financial variables. To overcome this fact, we propose the use of copulas to analyze the relation between domestic European indexes, conditional to the Euro Stoxx. The use of copulas allows us to model the dependence better than with elliptic distributions. We estimate conditional copulas using nonparametric methods to obtain the joint probability distribution function between indexes and examine the asymptotic properties of the nonparametric estimator. Moreover, we focus on the bandwidth choice of the conditional copula estimator and its sensitivity via a simulation study. This conditional copula method allows us to relate daily data with monthly data in a very simple manner, and therefore to relate stock indexes with macroeconomic variables such as inflation or GDP. We will measure the dependence and the conditional dependence using the Kendall's tau. We provide a statistic to test the significance of tau and its empirical distribution using jacknife.

C1758: **Vine copulas for multivariate time series modelling**
Presenter: **Esben Hoeg**, Aalborg University, Denmark
The so-called vine copulas are used to model multivariate time series. The methods are applied to a multidimensional time series problem within wind power generation. We use a time-varying copula mixture for the joint behavior of the individual univariate time series. The model is fitted to several years of German data, and captures the marginal behavior of the individual variables as well as the time variation in the dependence between the variables.

C1684: **Modeling the dependence structure of VIX and SP500**
Presenter: **Edit Rroji**, Universita' degli studi di Milano-Bicocca, Italy
Co-authors: Fabio Bellini, Lorenzo Mercuri
Copula-GARCH models are employed in the presence of multivariate time series as an alternative to standard multivariate GARCH models. Typically, univariate ARMA-GARCH models are fiited to the marginals while a dynamic copula is fitted to the innovations. We compare the fitting of different copula-GARCH models that describe the dependence structure of SP500 logreturns with the changes in the levels of Volatility Implied Index (VIX). Starting from findings in the financial literature, we require to candidate models to be able to reproduce two main features: strong negative dependence and radial asymmetry. We replicate the analysis in several subsequent time intervals in order to investigate the stability of results.

C1519: **Measuring exposure to dependence risk with random Bernstein copula scenarios**
Presenter: **Bertrand Tavin**, EMLYON Business School, France
The considered problem is that of measuring the exposure to dependence risk carried by a portfolio with an arbitrary number of two-asset derivative contracts. We develop a worst-case risk measure computed over a set of dependence scenarios within a divergence restricted region. The set of dependence scenarios corresponds to Bernstein copulas obtained by simulating random doubly stochastic matrices. We then devise a method to compute hedging positions when a limited number of hedging instruments are available for trading. In an empirical study we show how the proposed method can be used to reveal an exposure to dependence risk where usual sensitivity methods fail to reveal it. We also illustrate the ability of the proposed method to generate parsimonious hedging strategies in order to reduce the exposure to dependence risk of a given portfolio.

C1318: **Testing for structural breaks in factor copula models**
Presenter: **Florian Stark**, University of Cologne, Germany
Co-authors: Dominik Wied, Hans Manner
New fluctuation tests are proposed for detecting structural breaks in factor copula models and analyze the behavior under the null hypothesis of no change. In the model, the joint copula is given by the copula of random variables which arise from a factor model. This is particularly useful for analyzing data with high dimensions. Parameters are estimated with the simulated method of moments (SMM). Due to the discontinuity of the SMM objective function, it is not trivial to derive a functional limit theorem for the parameters. We analyze the behavior of the tests in Monte Carlo simulations and a real data application. In particular, it turns out that our test is more powerful than nonparametric tests for copula constancy in high dimensions.

CG243 Room MAL 541 CONTRIBUTIONS IN HIGH FREQUENCY FINANCIAL MODELLING Chair: Eduardo Rossi

C1670: **Statistical inference for price sluggishness**
Presenter: **Aleksey Kolokolov**, Goethe University Frankfurt, Germany
Co-authors: Davide Pirino, Giulia Livieri
Asset prices recorded at a high frequency are more sluggish than implied by the semi-martingale hypothesis. We propose a new general framework formalizing this phenomenon. We provide a limit theory for Multi-Idle-Time (an economic indicator for price idleness) and related quantities. This allows to quantify the level of idleness in an asset price adjustment and to test two different hypotheses. First, whether the extent of sluggishness is constant (and deterministic) or time-varying (and stochastic). Second, whether the sluggishness is persistent. The empirical application on US stocks provides the evidence that stock price flatness is both time-varying and persistent, especially during the crisis.

C1743: **Models for high-frequency trading volume data**
Presenter: **Eduardo Rossi**, University of Pavia, Italy
Co-authors: Paolo Santucci de Magistris
In finance theory prices are often supposed to follow an Ito semimartingale while no explicit assumptions are made on thedynamic evolution of trading volumes. The trading volume of a given security, i.e. the number of shares that change owners in a giveninterval of time, depend on the amount of new information available about the company. The intradaily trading volume series arecharacterized by large movements, so that the realizations are overdispersed. Models of high-frequency trading volume dynamics should becharacterized by large variance of the unconditional distributions. To this purpose, we propose a discrete-time model which has amultiplicative structure with time-varying parameters. We present the estimation and testing procedures and we validate the model throughan extensive empirical analysis.

C1599: **Overpersistence of oversupply: Measuring perceived persistence of oil shocks**
Presenter: **Veronika Selezneva**, CERGE-EI, Czech Republic
Co-authors: Stanislav Anatolyev, Sergei Seleznev
The oil market in the US is going through a major transformation due to the shale boom, creating temporary supply and demand imbalances along the way. We aim to estimate a perceived persistence of these shocks by market participants in the oil futures market. We use announcements about oil inventories to identify the imbalance shocks. We estimate an ARX-ARCHX-ACDX model of joint dynamics of returns, return volatilities and trading volumes around the announcements using high frequency data (5-second time intervals) on short and long maturity oil futures contracts. Long contracts are critical to distinguish shocks by persistence, as the effects of temporary shocks should vanish with maturity. Our model (i) handles illiquidity of long maturity contracts by explicitly accounting for trading inactivity; (ii) sheds light on the idiosyncratic trading behavior, capturing time varying trading intensity; and (iii) allows for structural changes in the dynamics and response to news over time, while also accounting for a unique sequential nature of oil stocks announcements. We find a complex and asymmetric response of the term structure curve to news, and document its evolution over time. Trading intensity differs substantially across contract maturities. We use our results to contribute to the debate on the nature of three recent major oil price shifts.

C1767: **Estimation of zero intelligence models by L1 data**
Presenter: **Martin Smid**, Institute of Information Theory and Automation, Czech Republic
A unit volume zero intelligence (ZI) model is defined and the distribution of its L1 process is recursively described. Further, a generalized ZI (GZI) model allowing non-unit market orders, shifts of quotes and general in-spread events is proposed and a formula for the conditional distribution of its quotes is given, together with a formula for price impact. For both the models, MLE estimators are formulated and shown to be consistent and asymptotically normal. Consequently, the estimators are applied to data of six US stocks from nine electronic markets. It is found that more complex variants of the models,despite being significant, do not give considerably better predictions than their simple versions with constant intensities.

C1804: **On the adjustment od international stock prices to an exchange rate shock**
Presenter: **Bahar Ghezelayagh**, University of East Anglia, United Kingdom
The intra-day GBP/USD exchange rates along with equity quotes for all British firms from New York (NYSE) and London (FTSE100 & FTSE250) during overlapping trading hours are analysed. The aim is to investigate where the price discovery happens and how stock prices adjust to an exchange rate shock.

Monday 18.12.2017 15:25 - 16:40 Parallel Session O – CFE-CMStatistics

EO047 Room Bloomsbury MODERN STATISTICAL METHODS FOR BIOMEDICAL AND CORRELATED DATA Chair: Zhiwei Zhang

E0669: **A quantitative concordance measure for comparing and combining treatment selection markers**
Presenter: **Zhiwei Zhang**, University of California at Riverside, United States
Co-authors: Shujie Ma, Lei Nie, Guoxing Soon

Motivated by an HIV example, the aim is to compare and combine treatment selection markers, which are essential to the notion of precision medicine. The current literature on precision medicine is focused on evaluating and optimizing treatment regimes, which can be obtained by dichotomizing treatment selection markers. In practice, treatment decisions are based not only on efficacy but also on safety, cost and individual preference, making it difficult to choose a single cutoff value for all patients in all settings. It is therefore desirable to have a statistical framework for comparing and combining treatment selection markers without dichotomization. We provide such a framework based on a quantitative concordance measure, which quantifies the extent to which higher marker values are predictive of larger treatment effects. For a given marker, the proposed concordance measure can be estimated from clinical trial data using a U-statistic, which can incorporate auxiliary covariate information through an augmentation term. For combining multiple markers, we propose to maximize the estimated concordance measure among a specified family of combination markers. A cross-validation procedure can be used to remove any re-substitution bias in assessing the quality of an optimized combination marker. The proposed methodology is applied to the HIV example and evaluated in simulation studies.

E1106: **Envelopes for censored quantile regression**
Presenter: **Yue Zhao**, KU Leuven, Belgium

Quantile regression has emerged as a powerful tool for survival analysis with censored data. We propose an efficient estimator for the coefficients in quantile regression with censored data using the envelope model. The envelope model uses dimension reduction techniques to identify material and immaterial components in the data, and forms the estimator of the regression coefficient based only on the material component, thus reducing the variability of the estimation. We will derive asymptotic properties of the proposed estimator and demonstrate its efficiency gains compared to the traditional estimator for the quantile regression with censored data. Recent advances in algorithms for the envelope model allow for efficient implementation of the proposed method. The strength of our proposed method is demonstrated via simulation studies.

E1359: **Hermite polynomial estimation of index models after a nonparametric transformation**
Presenter: **Yundong Tu**, Peking University, China

The purpose is to estimate transformation models with the index structure by using Hermite Polynomials. The estimation is performed using least squares, which is computational efficient for practical implementation. Asymptotic properties of the estimators of the index parameter, the unknown transformation function, the unknown link function are studied. Simulation studies show that the estimators enjoy nice finite sample properties and outperform the competitors available in the literature. The procedure is illustrated with a real data example.

EO451 Room Chancellor's Hall BAYESIAN MODEL SELECTION IN NONPARAMETRIC PROBLEMS Chair: Anirban Bhattacharya

E1024: **Scalable computation with shrinkage priors**
Presenter: **Anirban Bhattacharya**, Texas AM University, United States

The aim is to discuss some recent developments in scaling Markov chain Monte Carlo methods to Big (n, p) regression models, i.e., where the number of subjects and the dimensionality can both be large. We present an exact algorithm whose computational complexity scales linearly in the dimensionality and a randomized extension to deal with large n problems. Several applications to high-dimensional linear and logistic regression, dictionary learning, reduced-rank regression are illustrated.

E1035: **Hierarchical Gaussian process with local polynomial adjustments for big data regression**
Presenter: **Debdeep Pati**, Texas A&M University, United States
Co-authors: Anirban Bhattacharya, Hanning Li

In big data settings where the sample size is prohibitively large to invert kernel matrices in a Gaussian process regression, a popular approach is to use divide and conquer. The approach randomly partitions the data into a large number of sub-groups, fits a function to each group independently and averages these functions to obtain a global estimate. In order to induce coherence of the fitted function and to allow borrowing of information across multiple subgroups, we propose a hierarchical Gaussian process model on the group-specific functions. The common mean of the hierarchical Gaussian process is assumed a Gaussian process pivoted on a regular grid with local polynomial adjustments to correct bias. The covariance matrix of the latent Gaussian process has a "block-Toeplitz with Toeplitz-Blocks" structure in a bivariate setting. We propose a novel approach for sampling this latent Gaussian process by embedding this covariance matrix into a circulant matrix and using elliptical slice sampler to facilitate posterior computation. This results in a dramatic improvement of computational complexity compared to using a standard divide and conquer framework which assumes independence across the subgroups. The induced dependence provably results in a lower mean squared error compared to the standard divide and conquer framework. The approach is illustrated on a variety of synthetic examples and to a massive ecological dataset for predicting biomass.

E1453: **Horseshoe regularization for feature subset selection**
Presenter: **Jyotishka Datta**, University of Arkansas, United States
Co-authors: Nicholas Polson, Anindya Bhadra, Brandon T Willard

Feature subset selection arises in many high-dimensional applications of statistics, such as compressed sensing and genomics. The ℓ_0 penalty is ideal for this task, the caveat being it requires the NP-hard combinatorial evaluation of all models. A recent area of considerable interest is to develop efficient algorithms to fit models with a non-convex ℓ_γ penalty for $\gamma \in (0,1)$, which results in sparser models than the convex ℓ_1 or lasso penalty, but is harder to fit. We propose an alternative, termed the horseshoe regularization penalty for feature subset selection, and demonstrate its theoretical and computational advantages. The distinguishing feature from existing non-convex optimization approaches is a full probabilistic representation of the penalty as the negative of the logarithm of a suitable prior, which in turn enables an efficient expectation-maximization algorithm for optimization and MCMC for uncertainty quantification. In synthetic and real data, the resulting algorithm provides better statistical performance, and the computation requires a fraction of time of state of the art non-convex solvers.

EO727 Room Court STATISTICS FOR LARGE, SPATIAL DATASETS: ATMOSPHERIC SCIENCE APPLICATIONS Chair: Veronica Berrocal

E0383: **Modeling non-stationarity via multi-resolution basis functions and mixture priors**
Presenter: **Veronica Berrocal**, University of Michigan, United States
A typical challenge in air pollution population epidemiological studies is the lack of information on ambient exposure for most subjects. To circumvent this problem, and derive point-level estimates of air pollution, several methods have been proposed, including spatial statistical models that rely on the assumption of stationarity. This assumption might not be appropriate for PM2.5, a mixture of air pollutants that include both long-range contaminants and pollutants from more localized sources. To address this issue, building upon the M-RA model introduced recently, we express the spatial field as a linear combination of multi-resolution basis functions, and we provide the basis function weights with resolution-specific mixture priors. Simulation experiments demonstrate the ability of our model to detect regions of non-stationarity. Additionally, an application to daily average PM2.5 concentration indicates that: (i) the pattern of the spatial dependence of PM2.5 is non-homogeneous and (ii) our model outperforms ordinary kriging and is comparable to a previous non-stationary model in an out-of-sample prediction setting.

E0373: **Physically constrained spatiotemporal kriging of remotely sensed land surface temperature**
Presenter: **Matthew Heaton**, Brigham Young University, United States
Satellite remote-sensing is often used to collect important atmospheric and geophysical data that provide insight into spatial and temporal climate variability over large regions of the earth, at high spatial resolutions. Common issues surrounding such data include missing information in space due to cloud cover at the time of a satellite passing and large blocks of time for which measurements are not available due to the infrequent passing of polar-orbiting satellites. While many methods are available to predict missing data in space and time, in the case of land surface temperature (LST) data, these approaches generally ignore the temporal pattern called the diurnal cycle which physically constrains temperatures to peak in the early afternoon and reach a minimum just prior to sunrise. In order to construct a complete, physically justifiable remotely sensed dataset, we parameterize the diurnal cycle into a functional form with unknown spatiotemporal parameters. Using multiresolution basis functions, we estimate these unknown parameters from sparse satellite observations to obtain associated physically constrained predictions of land surface temperature. The methodology is demonstrated using a remote sensing dataset of LST in Houston, Texas USA, collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument, aboard NASAs polar-orbiting Aqua and Terra satellites.

E1070: **Compression and conditional emulation of climate model outputs**
Presenter: **Dorit Hammerling**, National Center for Atmospheric Research, United States
Numerical climate model simulations runs at high spatial and temporal resolutions generate massive quantities of data. As our computing capabilities continue to increase, storing all of the generated data is becoming a bottleneck, and thus is it important to develop methods for representing the full datasets by smaller compressed versions. We propose a statistical compression and decompression algorithm based on storing a set of summary statistics as well as a statistical model describing the conditional distribution of the full dataset given the summary statistics. The statistical model can be used to generate realizations representing the full dataset, along with characterizations of the uncertainties in the generated data. Thus, the methods are capable of both compression and conditional emulation of the climate models. Considerable attention is paid to accurately modeling the original dataset, particularly with regard to the inherent spatial nonstationarity in global temperature fields, and to determining the statistics to be stored, so that the variation in the original data can be closely captured, while allowing for fast decompression and conditional emulation on modest computers.

EO645 Room G11 COMPUTATIONAL INTENSIVE APPLICATIONS IN BIOSTATICS AND DNA EVIDENCE Chair: Joseph Gastwirth

E1222: **Pushing the boundaries for forensic DNA interpretation**
Presenter: **Therese Graversen**, University of Copenhagen, Denmark
Statistical interpretation of DNA from forensic evidence in crime cases may be computationally extremely demanding if the sample contains DNA from many people. While the forensic and statistical interpretation of the DNA sample concerns the DNA profiles of the people contributing to the sample, we may only observe the mixed signal of DNA components in the sample. The large discrete state space of the DNA profiles is the common root cause of a high computational complexity for all of the statistical models used for DNA interpretation in casework. The computational approach that we will show relies on well-established algorithms for computations in graphical models, one immediate advantage being that implementations are available in standard software. All computations are exact while still efficient enough to allow the interpretation of the more complex DNA samples occurring in practical casework. Many of the ideas are more generally applicable. One such example is to compute the likelihood in a model with discrete latent variables.

E1347: **A new approximate method for Y-STR haplotype probability assignment**
Presenter: **Jacob de Zoete**, Queen Mary University of London, United Kingdom
All current statistical models for the interpretation of Y-STR mixtures are hampered by a single point of failure-namely that there is no single agreed model for the assignment of Y haplotype probabilities. There are numerous competing models: the counting method, Brenners kappa method, the generalized Good method, the coalescent, the discrete Laplace, each with their strengths and weaknesses. We introduce a new approximate method, based on a Chow-Liu approximation of the dependency structure between loci. We will discuss the basis for its derivation and compare its performance to existing methods.

E1780: **The benefits and pitfalls of Bayes in forensic analysis**
Presenter: **Norman Fenton**, Queen Mary University of London, United Kingdom
Although the last forty years has seen considerable growth in the use of statistics in legal proceedings, it is primarily classical statistical methods rather than Bayesian methods that have been used. Yet the Bayesian approach avoids many of the problems of classical statistics and is also well suited to a broader range of problems. The aim is to address the potential and actual use of Bayes in the law - with special emphasis on forensic evidence - and explains the main reasons for its lack of impact on legal practice. These include misconceptions by the legal community about Bayes theorem, over-reliance on the use of the likelihood ratio and the lack of adoption of modern computational methods. We will argue that Bayesian Networks (BNs), which automatically produce the necessary Bayesian calculations, provide an opportunity to address most concerns about using Bayes in the law.

EO473 Room G3 LARGE-SCALE DATA WITH COMPLEX STRUCTURE Chair: Yixin Fang

E0228: **Estimation of subject-specific directed acyclic graphs with latent effects for discovering causal dependence**
Presenter: **Yuanjia Wang**, Columbia University, United States

Inferring causal relationship between variables from non-experimental data is a highly challenging goal, especially for large-scale data where estimation of directed acyclic graphs is NP-hard. Under the framework of structural equation models, or functional causal models, we represent joint distribution of variables in causal directed acyclic graphs (DAGs) as a set of structural equations, where directed edges connecting variables depend on subject-specific covariates and unobserved latent variables. The functional causal model framework allows constructing subject-specific DAGs, where the edges representing strength of network connectivity between variables decompose into a fixed effect term (average network effect given covariates), and a random effect term (unobserved residual network effect). Thus, our framework is a mixed effects DAG model. By pooling information across subjects under this model, we can estimate subject-specific network effects with a much better precision and assess heterogeneity of network effects with a better power. We theoretically prove identifiability of our model and propose a penalized likelihood based approach to handle high-dimensionality of the DAG model space, and a fast computational algorithm to achieve hard-thresholding of the edges.

E0634: **A first approach to real time sufficient dimension reduction**
Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom
Co-authors: Seung Jun Shin, Yuexiao Dong

A work in progress to create new effective and computationally efficient algorithms for real time sufficient dimension reduction (SDR) is presented. In an era where massive datasets are becoming the norm, we present a method where after an initial estimation based on our original data, when new data are collected the reduced subspace can be estimated using our proposed methods using only new data and completely ignoring the old data. We will demonstrate how this algorithm can be extended to sparse SDR.

E0460: **On scalable inference with stochastic gradient descent**
Presenter: **Yixin Fang**, New Jersey Institute of Technology, United States
Co-authors: Jinfeng Xu, Lei Yang

In many applications involving large data sets, stochastic gradient descent (SGD) provides a scalable way to compute parameter estimates and has gained increasing popularity due to its numerical convenience and memory efficiency. While the asymptotic properties of stochastic gradient descent-based estimators have been established decades ago, statistical inference such as interval estimation remains much unexplored. The traditional resampling method such as the bootstrap is not computationally feasible since it requires to repeatedly draw independent samples from the entire dataset. The plug-in method is not applicable when there are no explicit formulas for the covariance matrix of the estimator or the application comes from an online data setting where the sample arrives sequentially and it may not be necessary or realistic to store the entire dataset. We propose a scalable inferential procedure for stochastic gradient descent, which, upon the arrival of each observation, updates the SGD estimate as well as a large number of randomly perturbed SGD estimates. The proposed method is easy to implement in practice. We establish its theoretical properties in a general model setting which includes generalized linear models and quantile regression as special cases.

EO563 Room G4 ACCELERATED LIFE TESTING Chair: Maria Kateri

E0519: **Optimal designs for multiple-step-stress accelerated life testing experiments when some testing constraints are required**
Presenter: **Xiaojian Xu**, Brock University, Canada

Methods of designing accelerated life testing (ALT) experiments for the hazard rate prediction are investigated when a Cox proportional hazards (PH) model is utilized. We consider multiple step-stress ALT plans with time-censoring while certain practical testing requirements are needed. The maximum likelihood method is used for estimating the model parameters. The information matrix is derived, and the optimal stress levels and the optimal stress-changing times are determined under three optimality criteria: D-, A-, and Q-optimalities for the PH models with either a simple linear or a quadratic baseline hazard function. The efficiencies of our resulting optimal three-step-stress ALT plans are compared with their counterparts of optimal simple step-stress ALT plans. A practical simulation procedure is also provided and carried out to evaluate our resulting optimal three-step-stress ALT designs. Both asymptotic efficiency comparison and the simulation results have shown that the three-step-stress designs we have obtained with both an optimal middle stress level and two optimal stress changing times are more efficient than their competitors in terms of asymptotic variance, simulated bias, simulated variance, and simulated mean squared errors of the maximum likelihood estimator of the predicted hazard rate over a given period of time under normal design conditions.

E0631: **Adaptive step-stress accelerated life testing models**
Presenter: **Christian Kohl**, RWTH Aachen University, Germany
Co-authors: Maria Kateri

Under a step-stress accelerated life testing (SSALT) experiment the test items are exposed to stress levels that increase at intermediate time points of the experiment. In standard SSALT experiments the levels of the stresses applied at the various steps of the experiment are prespecified. An adaptive SSALT experiment is proposed, under which only the first stress level is fixed in advance while each subsequent stress level is chosen according to the number of failures on the immediately previous step. Assuming that the lifetimes of the testing items are exponentially distributed under each stress level, we derive the maximum likelihood estimators (MLEs) of the associated parameters and their exact distribution (conditional on their existence) under type-I and type-II censoring. Optimal test planning (i.e. optimal allocation of the stress change time points) in terms of C-, D- and A-optimality is considered for various scenarios. Several designs are compared via extensive simulation studies.

E0717: **A Cox model for component lifetimes with spatial interactions**
Presenter: **Christian Paroissin**, Universita de Pau et des Pays de l'Adour, France
Co-authors: Franck Corset, Mitra Fouladirad

By considering n components displayed on a structure (e.g., a steel plates), a Cox-type model for the hazard function is proposed which includes spatial dependency between components. The state (non-failed or failed) of each component is observed at some inspection times. From this data, we propose to use the SEM algorithm to estimate the parameters of the model. A study based on numerical simulation is then provided.

EO172 Room Gordon NON-PARAMETRIC ESTIMATION OF STATISTICAL MODELING Chair: Shanshan Ding

E1065: **Nonparametric estimation in a multi-armed bandit problem with covariates**
Presenter: **Wei Qian**, University of Delaware, United States
Co-authors: Yuhong Yang
The multi-armed bandit problem is a popular online decision-making problem with a wide range of modern applications in data science. Its classical setting consists of multiple slot machines with unknown expected payoffs, and the goal of the optimization game is to design a sequential arm allocation algorithm to maximize the total payoff. Motivated by promising applications in personalized medical and online web service, we consider a setting where the mean rewards of bandit machines are associated with covariates. With the key tradeoff between exploring new information and exploiting history information, we propose a kernel estimation based sequential allocation algorithm with randomization, and investigate its asymptotic and finite-time optimality under a nonparametric framework. In addition, since many nonparametric and parametric methods in supervised learning may be applied to estimating the mean reward functions, we integrate a model combining strategy into the allocation algorithm for adaptive performance. Simulations and real data evaluation are conducted to illustrate the algorithm performance and support the necessary consideration of covariates.

E1388: **A Dirichlet process mixture model for nonparametric Bayesian quantile regression**
Presenter: **Nan Lin**, Washington University in St. Louis, United States
A new nonparametric Bayesian approach to quantile regression is proposed by using a Dirichlet process mixture of logistic distributions (DPML). Many existing Bayesian quantile regression methods are based on parametric substitution of the error distribution by the asymmetric Laplace distribution, which is inconsistent with the typical nonparametric nature of quantile regression. The logistic distribution has a simple form in its quantile function and hence easily accommodates the quantile constraint. The proposed DPML model enjoys great model flexibility by mixing over both the location parameter and the scale parameter. We further established the posterior consistency of our proposed model and provided Markov chain Monte Carlo algorithms for posterior inference. The performance of our approaches is evaluated using simulated data and real data.

E1530: **A study on a general family of asymmetric distributions**
Presenter: **Md Rezaul Karim**, Katholieke Universiteit Leuven, Belgium
Co-authors: Irene Gijbels, Anneleen Verhasselt
A general family of asymmetric distributions is studied in which the location parameter is a specific quantile of the distribution (which is the main attraction of this family). We present the resulting special cases of asymmetric normal, asymmetric student-t and asymmetric logistic distributions and investigate their properties. We discuss parameter estimation procedures by using method of moments and maximum likelihood estimation. We briefly discuss advantages and disadvantages of the estimation methods, and establish their asymptotic behaviour. The proposed family of asymmetric distributions can be applied in a conditional setting, allowing to study regression quantiles in a likelihood framework. We consider both parametric and semi-parametric estimation in quantile regression settings.

EO063 Room Jessel ORDER RESTRICTIONS Chair: Satya PrakashSingh

E0298: **Some issues in the design of experiments with ordered experimental treatments**
Presenter: **Satya PrakashSingh**, University of Haifa, Israel
Co-authors: Ori Davidov
There are many situations where one expects an ordering among $K \geq 2$ experimental groups or treatments. Although there is a large body of literature dealing with the analysis under order restrictions, surprisingly very little work has been done in the context of the design of experiments. We provide some key observations and fundamental ideas which can be used as a guide for designing experiments when an ordering among the groups is known in advance. The power of the restricted likelihood ratio test is used as a design selection criterion. To address the dependency of the design criterion on the unknown model parameters, a maxi-min approach is incorporated. These maxi-min designs are then compared to some of the existing designs from the literature.

E1600: **Order-restricted inference in chronobiology**
Presenter: **Yolanda Larriba**, Universidty of Valladolid, Spain
Co-authors: Cristina Rueda, Miguel Fernandez
Biological processes, such as cell cycle, circadian clock or blood pressure, are governed by oscillatory systems consisting of numerous components that exhibit periodic patterns over time. Modelling these rhythms is a challenge in literature since usually the sampling density and the number of periods are low, and the underlying signals adopt a wide range of temporal patterns. Several authors proposed parametric functions of time, such as the sinusoidal function, to model these signals. However, these parametric functions might be too rigid for data derived from cell-cycle or circadian clock. These signals usually have a unique peak at time point U and a unique trough at time point L within each period, so that they monotonically increase up to U (when $L > U$) and then decrease up to L; before increasing again. The shape of these signals can be entirely described in the Euclidean space by mathematical inequalities among their components. The main novelty is the definition of circular signals using restrictions to model common signal shapes in biology. We will give a definition that allows us to state equivalent signal formulations both in the euclidean and in the circular spaces. Additionally, a general methodology is proposed to analyse rhythmicity including an efficient algorithm to compute the restricted maximum likelihood estimate, rhythmicity tests or an approach to estimate sampling order. Results are given for simulations and real data bases.

E1649: **The asymptotic distribution of the isotonic regression estimator over a general countable pre-ordered set**
Presenter: **Vladimir Pastukhov**, Lund University, Sweden
Co-authors: Dragi Anevski
The problem of isotonic regression over a general countable pre-ordered set is studied. We prove that the limiting distribution of the isotonised estimator is given by the concatenation of the separate isotonic regressions of the restrictions of an underlying estimators asymptotic distribution to the comparable level sets of the underlying estimators probability limit. Also, we show that the isotonisation preserves the rate of convergence of the underlying estimator. We apply the results to the problems of estimation of a bimonotone regression function and estimation of a bimonotone probability mass function.

EO250 Room Montague HIGH-DIMENSIONAL/FUNCTIONAL DATA ANALYSIS WITH BIOLOGICAL APPLICATIONS Chair: Juhyun Park

E1078: **A Bayesian model for drug response estimation and biomarker testing using Gaussian processes**
Presenter: **Frank Dondelinger**, Lancaster University, United Kingdom

Large-scale drug response assays on cancer cell lines have in recent years produced a wealth of data that benefits cancer treatment, drug development and biomarker discovery. We present a novel Bayesian approach for modelling dose-response curves using Gaussian processes (GPs). Our model has several advantages over the traditional sigmoid curve: 1) it is non-parametric and allows us to fit a variety of responses; 2) it allows for a hierarchical Bayesian setup with information sharing across different curves; and 3) we automatically obtain a measure of the uncertainty of our curve fits from the variance of the Gaussian process. We extend the model with a Bayesian biomarker testing framework that allows us to test for a difference in the proportion of responsive curves in mutated versus wild type cell lines. We test the model on cell line drug response data from the Cancer Genome Project, and demonstrate that the Gaussian process model shows greater robustness to outliers and to unusual response patterns. The Bayesian testing model successfully identifies known biomarkers, and is able to leverage information about the complete dose-response curve, rather than relying on summary measures.

E0740: **ABC inference for a spatio-temporal autologistic model**
Presenter: **Anne Gegout-Petit**, Universite de Lorraine-IECL Inria BIGS, France
Co-authors: Shuxian Li

A centered autologistic model is proposed to model the occurrence of an illness on a regular lattice. It allows us to test spatial dependencies and effect of covariates. In the framework of Markov chain of Markov field, we first prove the existence of the spatial law of the field at time t given the covariates and the past of the process. We show the utility of a good spatio-temporal centering for the correct modelling and interpretation of the effect of the neighbours on the occurrence. It preserves the large-scale structure of the phenomena. About inference, we show trough simulations, the good performance of the expectation-maximization (EM) pseudo-likelihood estimation. We study also the possibility to use Approximate Bayesian Computation (ABC) for model selection of the structure of neighbourhood.

E0822: **Hierarchical functional clustering using equivalence test with application to perfusion imaging**
Presenter: **Yves Rozenholc**, University Paris Descartes, France
Co-authors: Fuchen Liu, Charles A Cuenod

Perfusion imaging allows non invasive access to tissue micro-vascularization. Promising tool to build imaging biomarkers, it suffers from low SNR, improved by averaging homogeneous functional information in large regions of interest. We propose a new automatic segmentation of such image sequence into functionally homogeneous regions. At its core, HiSET (Hierarchical Segmentation using Equivalence Test) aims to cluster functional signals discretely observed with noise on a finite metric space. Assuming independent fixed Gaussian noise, HiSET uses p-values of a multiple equivalence test as dissimilarity measure. It consists of two steps only varying through the neighborhood structure. The first benefit from local regularities on the metric space to control the complexity, the second recovers (spatially) disconnected homogeneous structures at a larger scale. Given a maximal expected homogeneity discrepancy δ, both steps stop automatically through a control of the type I error, providing an adaptive segmentation. Tuning parameter δ can be interpreted as a multi-resolution diameter around functional patterns recover by the segmentation. When the landscape is functionally piecewise constant with well separated functional features, HiSET is proven to retrieve the exact partition with high probability when the number of observation times is large enough. HiSET outperforms state-of-the-art clustering methods for perfusion imaging sequences.

EO725 Room Senate ROBUST METHODS FOR COMPLEX DATA Chair: Ana Belen Ramos-Guajardo

E0462: **Two-sample dispersion tests for interval-valued data**
Presenter: **Przemyslaw Grzegorzewski**, Warsaw University of Technology, Poland

Two-sample tests for dispersion belong to a basic toolbox of statistical procedures. The goal is to generalize a suitable test for interval-valued data perceived both from the epistemic and ontic perspective. Usually to verify whether two populations differ in scale one compares two population variances or standard deviations. Unfortunately, a generalization of such statistical procedures into the interval-valued framework may cause considerable computational problems, especially if a sample is large enough. Indeed, a sample variance computation for the epistemic intervals is NP-hard task. Moreover, most of the tests for comparing variances assume that the underlying population is normally distributed. Therefore, to avoid problems in verifying assumptions on the underlying distribution we consider nonparametric tests for dispersion which are not based on variances. They are not only distribution-free, but fortunately, unlike typical generalizations of statistical procedures into the interval-valued framework, our generalized tests entail very low computational costs.

E0935: **Generalization of the Mahalanobis distance for fuzzy data: An application to robust fuzzy clustering**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Maria Brigida Ferraro

Most of the distances used in case of fuzzy data are based on the well-known Euclidean distance. A fuzzy set can be characterized by centers and spreads and the distances between fuzzy sets are essentially defined as a weighted sum of the squared Euclidean distances between the centers and the spreads. In the multivariate case the Euclidean distance does not take into account the correlation structure between variables. For this reason, one possibility to avoid this drawback is to consider the Mahalanobis distance since it involves the covariance matrix between the variables. A generalization of that distance in the fuzzy framework is proposed. It is shown to be very useful in different contexts as, for instance, in the robust fuzzy clustering approach. Non-spherical clusters are not generally recognized by means of Euclidean-type distances whereas they are shown to be recognized if the generalized Mahalanobis distance is taken into account. Some theoretical properties are addressed and clustering applications are reported in order to check its adequacy.

E1597: **Estimation of the relative scale of trees from their Harris path**
Presenter: **Romain Azais**, INRIA Nancy, France
Co-authors: Alexandre Genadot, Benoit Henry

Hierarchical data naturally appear in various application fields, e.g., plant modeling and XML file analysis. The aim is to propose a new method to estimate the relative scale of ordered trees relying on a theoretical study for Galton-Watson trees conditioned on their number of nodes. Our approach is based on the adequacy of the Harris path of the tree with an expected function that describes the average shape of the tree. We prove the consistency of our estimators for conditioned Galton-Watson trees and compare them with a previous approach of the literature. Numerical investigations show the good behavior of our procedure on finite-sample sizes as well as from missing or noisy data. An application to the analysis of revisions of Wikipedia articles is also considered and states the robustness of our approach.

EO463 Room Woburn QUANTILE ESTIMATION AND REGRESSION Chair: Anneleen Verhasselt

E0416: **An adapted loss function for censored quantile regression**
Presenter: **Anouar El Ghouch**, The University catholique de Louvain, Belgium
Co-authors: Ingrid Van Keilegom, Mickael De Backer
A novel approach is studied for the estimation of quantiles when facing potential right censoring of the responses. Contrary to the existing literature on the subject, the adopted strategy is to tackle censoring at the very level of the loss function usually employed for the computation of quantiles, the so-called check function. For interpretation purposes, a simple comparison with the latter reveals how censoring is accounted for in the newly proposed loss function. Subsequently, when considering the inclusion of covariates for conditional quantile estimation, by defining a new general loss function, the proposed methodology opens the gate to numerous parametric, semiparametric and nonparametric modelling techniques. We consider the well studied linear regression under the usual assumption of conditional independence between the true response and the censoring variable. For practical minimization of the studied loss function, we also provide a simple algorithmic procedure shown to yield satisfactory results for the proposed estimator with respect to the existing literature in an extensive simulation study. From a more theoretical prospect, consistency of the estimator for linear regression is obtained using very recent results on non-smooth semiparametric estimation equations with an infinite-dimensional nuisance parameter, while numerical examples illustrate the adequateness of a simple bootstrap procedure for inferential purposes.

E0452: **Quantile regression estimation in varying coefficient models**
Presenter: **Irene Gijbels**, Katholieke Universiteit Leuven, Belgium
Quantile regression is an important tool for describing the characteristics of conditional distributions. In both, mean and quantile regression, flexible models are often needed to capture the complexity of the underlying stochastic phenomenon. Among flexible models encountered in a multivariate covariate regression setting are varying coefficient models. These models introduce additional flexibility as compared to multiple linear regression models by allowing the coefficients to vary with, for example, another covariate. Coefficients are no longer real parameters but unknown functions. We focus on estimating conditional quantile functions in varying coefficient models. As a major tool we use B-spline approximations for the unknown coefficient functions. We distinguish between homoscedastic and heteroscedastic modeling, and discuss also estimation of the variability function in the latter setting. We summarize the theoretical results obtained and briefly discuss practical implementation issues. The finite-sample performances of the estimation methods have been investigated via simulation studies and their practical use has been illustrated in real data analysis.

E0483: **Testing in heteroscedastic quantile varying coefficient models**
Presenter: **Anneleen Verhasselt**, Hasselt Univeristy, Belgium
Co-authors: Irene Gijbels, Mohammed Abdulkerim Ibrahim
Quantile regression in varying coefficient models is considered. Varying coefficient models are a flexible generalization of linear regresion models, in the sense that the coefficients are functions (of a covariate) instead of constants. Varying coefficient models have been used for modeling longitudinal data, where the coefficients are then allowed to vary with time. The coefficient functions are estimated with splines. We focus on developing a testing procedure for the shape (constant, monotone, convex) of the coefficient functions. Furthermore, we allow the error structure to be heteroscedastic and propose a likelihood-ratio-type testing procedure for comparing the variability functions. The performance of the procedures is investigated via simulation studies and their use is illustrated on data examples.

EO465 Room SH349 CLUSTERING/CLASSIFICATION AND MIXTURES II Chair: Geoffrey McLachlan

E0777: **Random projection ensemble clustering**
Presenter: **Angela Montanari**, Alma mater studiorum-Universita di Bologna, Italy
Co-authors: Francesca Fortunato, Laura Anderlucci
Random projections (RPs) have shown to provide promising results for high-dimensional classification. The RP ensemble classifier, in fact, overcomes the inherent instability of a single RP by using a "selection and ensemble aggregation" routine. The general idea of RP ensemble is extended to high-dimensional clustering purposes. Specifically, following the original procedure, the best projection according to a specific clustering quality measure is chosen within each of B_1 distinct blocks of B_2 RPs. Then, the final partition is obtained by aggregating, via consensus, results of applying a model-based clustering algorithm to the selected projections. The performances of the method are assessed both on a set of real and simulated data.

E0350: **Exploration of the variability of variable selection based on distances between bootstrap sample results**
Presenter: **Christian Hennig**, UCL, United Kingdom
Co-authors: Willi Sauerbrei
It is well known that variable selection in multiple regression can be unstable and that the model uncertainty can be considerable. The model uncertainty can be quantified and explored by bootstrap resampling. We will present approaches that use the results of bootstrap replications of the variable selection process to obtain more detailed information about the data. Analyses will be based on distances between the results of the analyses of different bootstrap samples. The distances are used to map the bootstrap results by mutidimensional scaling and to cluster them. Clusters are of interest because they could point to substantially different interpretations of the data that could arise from different selections of variables supported by different bootstrap samples. These and further issues will be illustrated by some data examples including a study on ozone effects in children.

E0425: **Personalized cancer-specific integrated network estimation**
Presenter: **Kim-Anh Do**, University of Texas MD Anderson Cancer Center, United States
Co-authors: Veerabhadran Baladandayuthapani, Min Jin Ha
The aim is to propose personalized cancer-specific integrated network estimation (PRECISE), a general framework for integrating existing interaction databases, data-driven de novo causal structures, and upstream molecular profiling data to estimate cancer-specific integrated networks, infer patient-specific networks and elicit interpretable pathway-level signatures. We develop a Bayesian regression model for protein-protein interactions that integrates with known pathway annotations and protein-protein interactions. In the first step of PRECISE, a data-driven protein causal network is estimated and combined with the prior information. Other upstream molecular data are then integrated in three sequential steps to produce cancer-specific, patient-specific networks and pathway scores, which are subsequently used for tumor subtype classification and clinical outcome prediction. Using the pan-cancer functional proteomic data on 32 cancer types from The Cancer Genome Atlas, we demonstrate the utility of PRECISE in inferring commonalities and differences in network biology across tumor lineages and in using patient-specific pathway-based signatures for robust tumor stratification and prediction.

EC702 Room G5 CONTRIBUTIONS TO EXTREME VALUE THEORY AND APPLICATIONS Chair: Marie Kratz

E0283: **Generalized means in statistical extreme value theory**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal
Co-authors: Frederico Caeiro, Manuela Neves, Helena Penalva
The focus is on Holder's and Lehmer's generalized mean-of-order-p of adequate statistics based on the k upper ordered observations associated with a stationary weakly dependent sample from a parent $F(\cdot)$. We are interested in the estimation of the extreme value index (EVI), the primary parameter in statistical extreme value theory (EVT) and related parameters, like the value-at-risk and the expected shortfall. The asymptotic behavior of the aforementioned classes of EVI-estimators enables their asymptotic comparison at optimal levels (k,p), in the sense of minimal mean square error. In the Monte-Carlo simulation study we further include the associated location-invariant PORT EVI-estimators, with PORT standing for peaks over random threshold.

E1695: **Asymptotic error bound approximation of threshold exceedance probabilities for non-stationary random fields**
Presenter: **Jose Luis Romero**, University of Granada, Spain
Co-authors: Jose Miguel Angulo
Evaluation of different forms of threshold exceedance probabilities, for large thresholds in general scenarios, is one of the most relevant problems arising in the assessment of extremal behaviour in real phenomena. Euler-Poincare characteristic (EPC) is a structural property of excursion sets, intrinsically related to these probabilities. There is a vast literature related to providing error bound approximations between excursion probabilities and EPC under Gaussian and/or stationary assumptions for the underlying random field model, with increasing interest in more general scenarios. Spatial deformation is used in fields such as image analysis, environmental studies, etc. as an approach to represent heterogeneities of a reference stationary random field. Local smoothing by means of a kernel-based blurring transformation in terms of a convolution operator is also used in real applications. From both transformations, significant classes of non-stationary random fields arise. There exist kernel-based regularizing sequences providing different forms of convergence to the random field reference model, by means of the derived blurred sequence. Asymptotic error bound approximations for non-regular random fields obtained under spatial deformation and blurring transformations of the reference model are extended. Effects of regularizing sequences such as, the so-called, Mollifiers and Gaussian kernels, is analyzed and illustrated by simulation for different scenarios.

E1543: **Consistent asymptotically normal estimators for stable distributions**
Presenter: **Annika Krutto**, University of Tartu, Estonia
Two methods for estimating parameters in general stable distributions are proposed. One allows point estimators for the scale and shift, another for the four parameters of stable distributions. Methods involve empirical cumulant function. The main advantage of the methods is computational simplicity. Moreover, both provide consistent estimators. Asymptotic normal distributions for the proposed cumulant estimators are specified. The efficiency of the method relies on two arbitrary real points. The points selection problem is treated and suggestions stipulated. Comprehensive simulations are performed. Comparison to other estimation procedures is presented.

CO595 Room MAL B18 TESTING SEMI-NONPARAMETRIC HYPOTHESES Chair: Tatiana Komarova

C0233: **Nonparametric significance testing in measurement error models**
Presenter: **Luke Taylor**, London School of Economics, United Kingdom
A nonparametric significance test for regression models with measurement error in the regressors is developed. To the best of our knowledge, this is the first test of its kind. We use a 'semi-smoothing' approach with nonparametric deconvolution estimators and show that our test is able to overcome the slow rates of convergence associated with such estimators. In particular, our test is able to detect local alternatives at the $\sqrt{n}$-rate. We derive the asymptotic distribution under weakly dependent data and provide a bootstrap procedure. We also highlight the finite sample performance of the test through a Monte Carlo study. Finally, we discuss two empirical applications. The first considers the effect of cognitive ability on a range of socio-economic variables: income, life satisfaction, health and risk aversion. The second uses time series data to investigate whether future inflation expectations are able to stimulate current consumption; an important policy question when nominal interest rates approach the zero lower bound.

C1246: **A simple test for monotonicity and monotonicity-related properties**
Presenter: **Tatiana Komarova**, London School of Economics and Political Science, United Kingdom
Co-authors: Javier Hidalgo
A test for monotonicity in a nonparametric framework is developed by using partial sums empirical process. We show that the test has suitable asymptotic properties. In particular we show after appropriate transformation the asymptotic distribution is a functional of a standard Brownian motion, so that critical values are available. However, due to the possible poor approximation of the asymptotic critical values to the finite sample ones, we also describe a valid bootstrap algorithm. We show how methodology can be extended to test for other properties of the regression function such as convexity, concavity, absolute monotonicity and U-shape. We outline how this can extended to a framework when other covariates are present and no monotonicity-related properties are imposed on those. We also establish how monotonicity can be tested in the situation of endogeneity if there is a strong instrument available.

C0922: **Jackknife, small bandwidth and high-dimensional asymptotics**
Presenter: **Yukitoshi Matsushita**, Tokyo Institute of Technology, Japan
Co-authors: Taisuke Otsu
A jackknife based inference procedure for semiparametric models is proposed. It is shown that a jackknife empirical likelihood statistic is asymptotically pivotal under standard asymptotics, while it is not under non-standard asymptotics allowing for small bandwidths or many covariates. A modified jackknife empirical likelihood is proposed that is fully automatic and works for under both standard and non-standard asymptotics. Our findings are applied to three non-standard settings: small bandwidth asymptotics for semiparametric density-weighted average derivatives, many/weak IV asymptotics for IV models, and many covariates asymptotics for linear regression models.

CO664 Room MAL B20 ECONOMETRICS WITH MIXED FREQUENCY AND AGGREGATED DATA Chair: Michael Thornton

C1051: **Discretization of the Bergstrom-Nowman macroeconomic model**
Presenter: **Roderick McCrorie**, University of St Andrews, United Kingdom
The discretization of the Bergstrom-Nowman macroeconomic model is discussed. This model represents the state-of-the-art in continuous-time macroeconometric modeling. Its estimation is facilitated by the derivation of an exact discrete analog of a linear approximation of the model about sample means. We focus on providing a compact form for the covariance matrix of the disturbance vector of this model. Each block of the covariance matrix is shown to be expressible compactly in terms of triple integrals involving the matrix exponential, which can be computed by combining submatrices of the exponential of a certain block triangular matrix. This offers dramatically simplified expressions for the Gaussian likelihood function of the model. In passing, some new results are presented on the aliasing identification problem and on the derivation of the exact discrete analog of certain, more general continuous time stochastic models.

C0862: **Representations of linear continuous time models with mixed frequency data**
Presenter: **Michael Thornton**, University of York, United Kingdom
Methods to derive an exact representation for observations generated at mixed frequencies by an underlying continuous time ARMA (CARMA) process are derived. The data may be stocks, flows or a mixture of the two. The method involves reconstructing a discrete time linear process, with first and second properties identical to the underlying model, from a state space model and avoids treating mixed frequency data as a problem of missing or aggregated observations.

C1412: **In-fill asymptotic theory for structural breakpoint in autoregression: A unified theory**
Presenter: **Xiaohu Wang**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Liang Jiang, Jun Yu
The exact distribution of the maximum likelihood estimator of structural break points in the Ornstein Uhlenbeck process is studied when a continuous record is available. The exact distribution is asymmetric, tri-modal, dependent on the initial condition. These three properties are also found in the finite sample distribution of the least squares (LS) estimator of structural break point in autoregressive (AR) models. Motivated by these observations, an in-fill asymptotic theory is developed for the LS estimator of structural break point in the AR(1) coefficient. The in-fill asymptotic distribution is also asymmetric, trimodal, dependent on the initial condition, and delivers excellent approximations to the finite sample distribution. Unlike the long span asymptotic theory, which depends on the underlying AR root and hence is tailor made but is only available in a rather limited number of cases, the in-fill asymptotic theory is continuous in the underlying roots. Monte Carlo studies show that the in fill asymptotic theory performs better than the long-span asymptotic theory for cases where the long span theory is available and performs very well for cases where no long-span theory is available.

CO676 Room MAL B35 INFERENCE IN COMPLEX DYNAMIC MODELS Chair: Alessandra Luati

C1234: **Testing the hypothesis of enhanced design in fast fashion industry using internet as a source of data**
Presenter: **Marzia Freo**, University of Bologna, Italy
Co-authors: Federico Crescenzi, Alessandra Luati
The aim is to assess the extent to which enhanced design is used in combination with quick response in the fashion system, following previous results. Using internet as source of data, we scrap web data from online sites of traditional and innovative fast fashion firms and propose decision rules to define when firms adopt enhanced design and/or quick response strategies. We then evaluate how firms implement the two policies and if these are used as complements or substitutes. Finally, we empirically test some implications of alternative strategies.

C1146: **Inference in the Duffing system with a sequential ABC-UKF algorithm**
Presenter: **Michela Eugenia Pasetto**, University of Bologna, Italy
Co-authors: Umberto Noe, Alessandra Luati, Dirk Husmeier
An algorithm is developed to infer the parameters of a nonlinear chaotic system of differential equations (Duffing oscillator) based on the Unscented Kalman Filter (UKF). We have found that the overall inference performance of the UKF and its convergence critically depend on the location of the so-called sigma points and the initialization of the algorithm. To address these limitations, we propose a novel algorithm called Sequential ABC-UKF. First, we apply Approximate Bayesian Computation (ABC) with Sequential Monte Carlo (SMC) to provide a starting value in parameter space for informed initialization of the UKF. Second, we optimize the sigma point locations, comparing two alternative schemes: Bayesian optimization versus exhaustive discrete grid search. We demonstrate the effectiveness of the proposed method with a simulation study and real data analysis.

C1434: **Fast inference in expensive computational models**
Presenter: **Umberto Noe**, University of Glasgow, United Kingdom
Co-authors: Dirk Husmeier, Maurizio Filippone, Nicholas Hill, Weiwei Chen
Inference in expensive computational models, involving the numerical solution of a system of Partial Differential Equations (PDEs), is discussed. These models are typically referred to as "black-box" functions and could arise from Biomechanical, Engineering or Financial problems. They are not suitable for MCMC or standard likelihood based inference due to the high computational resources needed for a single output and the non-availability of gradient information. We present an extension of the Efficient Global Optimization (EGO) algorithm that allows for hidden constraints. The latter often arise in practice when a simulation returns no output at a point that violates the model assumptions.We illustrate the improved algorithm on a PDE model of the double sided Human Pulmonary Circulation. The reduction in the computational cost, compared to standard likelihood-based methods, is by two orders of magnitude.

CO214 Room MAL 402 MANAGEMENT OF EXPECTATIONS IN TURBULENT TIMES Chair: Maritta Paloviita

C0733: **De-anchoring of inflation expectations under learning and heterogeneity**
Presenter: **Alberto Locarno**, Banca d'Italia, Italy
Co-authors: Fabio Busetti, Davide Delle Monache, Gerali Andrea
The purpose is to study how a prolonged period of subdued price developments may induce a de-anchoring of inflation expectations from the central banks objective. This is shown within a framework where agents form expectations using adaptive learning, choosing among a set of alternative forecasting models. Expectations are therefore non rational and heterogeneous: some agents are naive and form inflation expectations only on the basis of past data; others instead trust the central bank's objective and use models that better capture the working of the economy. The analysis is accompanied by empirical evidence on the properties of inflation expectations in the euro area. Our results suggest that monetary policy may lose effectiveness if delayed too much, as expectations are allowed to drift away from target for too long. Optimal monetary policies tend to be more hawkish than under rational expectations: the central bank must react more strongly to deviation of inflation from targets, regardless of whether inflation is too high or too low.

C1685: **Formation of inflation expectations in turbulent times: On ECB expectations' management of professional forecasters**
Presenter: **Tomasz Lyziak**, National Bank of Poland, Poland
Co-authors: Maritta Paloviita
The aim is to study formation of inflation expectations in the euro area. We first analyse forecast accuracy of ECB inflation projections relative to private sector forecasts. Then, using the ECB Survey of Professional Forecasters (ECB SPF), we estimate a model integrating two theoretical concepts: the hybrid model of expectations, including rational and static expectations, and sticky-information model. We examine whether ECB inflation projections are still important in expectations formation once the impact of forward-lookingness of economic agents has been taken into account. We also derive implicit inflation targets and assess their consistency with the official ECB inflation target. Our analysis indicates that recent turbulent times have contributed to changes in expectations formation, as the importance of backward-looking mechanisms has decreased, while the importance of the perceived inflation target has increased. We also find that the perceived inflation target has remained broadly consistent with the official ECB target in the medium-term. However, its downward trend suggests some risks of de-anchoring of inflation expectations. The importance of ECB inflation projections for medium-term SPF forecasts has increased over time, but the magnitude of this effect is rather small. However, SPF inflation forecasts remain consistent with the ECB communication, being either close to ECB projections or between ECB projections and the inflation target.

C1793: **Semantic similarity in central bank communication and market volatility**
Presenter: **Jonathan Talmi**, Bank of Canada, Canada
Co-authors: Michael Ehrmann
Press releases announcing and explaining monetary policy decisions play a critical role in the communication strategy of central banks. Due to their market-moving potential, it is particularly important how they are drafted. Often, central banks start from the previous statement, and update the earlier text at the margin. This makes it straightforward to compare statements and see how the central banks thinking has evolved; however, more substantial changes, which will eventually be required, might then be harder to understand. Using variation in the drafting process at the Bank of Canada, the extent to which similarity in central bank statements matters for the reception of their content in financial markets is studied. It is shown that similar press releases generate less market volatility, but that more substantial textual changes after a sequence of very similar statements lead to much larger volatility.

CO122 Room MAL 414 FINANCIAL TIME SERIES MODELLING Chair: Zudi Lu

C0600: **Robustness and vulnerability of networks with dynamical dependency groups**
Presenter: **Lei Wang**, Beihang University, China
The dependency property and self-recovery of failure nodes both have great effects on the robustness of networks during the cascading process. Existing investigations focused mainly on the failure mechanism of static dependency groups without considering the time-dependency of inter-dependent nodes and the recovery mechanism in reality. We present an evolving network model consisting of failure mechanisms and a recovery mechanism to explore network robustness, where the dependency relations among nodes vary over time. Based on generating function techniques, we provide an analytical framework for random networks with arbitrary degree distribution. In particular, we theoretically find that an abrupt percolation transition exists corresponding to the dynamical dependency groups for a wide range of topologies after initial random removal. Moreover, when the abrupt transition point is above the failure threshold of dependency groups, the evolving network with the larger dependency groups is more vulnerable; when below it, the larger dependency groups make the network more robust. Numerical simulations employing the Erdos-Renyi network and Barabsi-Albert scale free network are performed to validate our theoretical results.

C0581: **An extended MRR Model for transaction-level analysis of high frequency trading processes**
Presenter: **Qiang Zhang**, Beijing Universtiy of Chemical Technology, China
A nonparametric test is employed, showing that the Markov property of the trade indicator variables, which is a key assumption in the MRR model, is rejected in most of trading days. Based on the spread decomposed structure, an extended MRR model is proposed as an extension. Empirical results show that the information lag plays an important role and the difference of the adverse selection risk parameter between two models is significant. Further, our analysis suggests that the information lag parameter can be a useful measure of the average speed at which the information integrates into the price.

C1470: **Metallgesellschaft's hedging revisited: A bootstrap reality check**
Presenter: **Tyler Brough**, Utah State University, United States
The derivatives trading episode of Metallgesellschaft (MGRM) is now more than two decades old. The case still captures great interest in research and teaching as it pertains to lessons learned from using derivatives for risk management. The episode spurred much early academic research, much of it contradictory. It has been previously defended MGRM's hedging practices as essentially sound. At the same time other authors provide evidence that MGRM's hedging was excessively speculative relative to a minimum-variance benchmark. We bridge the gap between this literature by providing more direct evidence. We do this by applying a nonlinear vector error correction model with GARCH to establish the dynamic properties of spot and futures prices. Having established these properties and the nature of the cointegrating relationship, we then employ White's bootstrap Reality Check to directly compare Pirrong's optimal BAG minimum-variance hedge ratio and MGRM's one-for-one hedging ratio strategy. Further, we utilize the simulation to address the liquidity risks faced by MGRM and how they might have been better managed. We find evidence that MGRM's hedging was sound but that by utilizing a synthetic funding strategy they could have improved their liquidity risk substantially.

CO118 Room MAL 421 LONG MEMORY Chair: Josu Arteche

C0264: **Origins of spurious long memory**
Presenter: **Philipp Sibbertsen**, University of Hannover, Germany
Co-authors: Christian Leschinski
A large class of structural change processes that generate spurious long memory are considered. Among others, this class encompasses structural breaks as well as random level shift processes and smooth trends. The properties of these processes are studied based on a simple representation of their discrete Fourier transform. We find that, under very general conditions, all of the models nested in this class generate poles in the periodogram at the zero frequency. These are of order $O(T)$, instead of the usual $O(T^{2d})$ for long memory processes and $O(T^2)$ for a random walk. This order arises whenever both the mean changes and sample fractions at which they occur are non-degenerate, asymptotically.

C0362: **Multiple local Whittle estimation of long memory**
Presenter: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain
The estimation of the memory parameter in long memory series has attracted great attention from the last few years of the 20th century onwards. Most research has focused on standard long memory at frequency zero, where several semiparametric estimators have been proposed for stationary, non-stationary and non-invertible series. However, much less attention has been paid to the existence of seasonal or cyclical strong persistence,

where several spectral poles can appear at non-zero frequencies. In fact, the existing proposals only cover the stationary and invertible case. Based on the Exact Local Whittle estimation for standard long memory, we propose a semiparametric estimation technique for jointly estimating all the memory parameters in seasonal and cyclical long memory time series. Consistency and asymptotic normality are proved for stationary, non-stationary and non-invertible series, allowing for straightforward standard inference of interesting hypotheses such as the existence of unit roots at some or all seasonal frequencies. The fine finite sample performance of our procedure is analysed via Monte Carlo.

C0629: **Estimation for dynamic panel data with individual effects**
Presenter: **Carlos Velasco**, Universidad Carlos III de Madrid, Spain
Co-authors: Peter Robinson
Statistical inference on parametric models for panel data is discussed. The models feature dynamics of a general nature, individual effects and possible explanatory variables. The focus is on large-cross-section inference on Gaussian pseudo maximum likelihood ratio estimates with temporal dimension kept fixed, partially complementing and extending recent work of the authors. We focus on a particular kind of initial condition but go on to discuss implications of alternative initial conditions. Some possible further developments are briefly reviewed.

CO481 Room MAL 532 MONETARY POLICY AND FINANCIAL CONDITIONS Chair: Luca Benzoni

C0718: **Delphic and Odyssean monetary policy shocks: Evidence from the euro-area**
Presenter: **Filippo Ferroni**, Federal Reserve Bank of Chicago, United States
The impact of the ECB announcements on the market-based expectations of interest rates and of inflation rates is studied. We find that the impact of the ECB announcements on inflation expectations has changed over the last fifteen years. In particular, while in the central part of our sample the ECB announcements were read as a signal about the economic conditions (i.e. Delphic component), in latest episodes they have been interpreted as a commitment device on future monetary policy accommodation (i.e. Odyssean component). We propose an approach to separately identify the Delphic and Odyssean component of the ECB monetary policy announcements and we measure their dynamic impact on the economy.

C0397: **Term premium, credit risk premium, and monetary policy**
Presenter: **Andrea Ajello**, Board of Governors of the Federal Reserve System, United States
The aim is to build and calibrate a New-Keynesian DSGE model with Epstein-Zin preferences and financial frictions in the shape of multi-period nominal defaultable debt, to fit U.S. data moments. We solve the model using higher-order perturbations and show that credit frictions can significantly increase the size and volatility of the nominal and real Treasury term premium through the interaction of preferences sensitive to long-run risk, and amplification of the economy's response to TFP shocks. Our analysis suggests that introducing multi-period defaultable debt contracts helps fit the cyclical properties of macroeconomic and financial variables, including credit spreads, credit risk premia and leverage ratios together with the main features of the *default-free* term structure of interest rates. Model simulations show that unexpected monetary policy shocks have small effects on term and credit-risk premia dynamics, while the systematic component of monetary policy has sizable implications for the average and volatility of risk compensation. In particular, monetary policy that responds more to inflation fluctuations relative to output reduces the average and the volatility of nominal term premia, while increasing the average and volatility of credit risk premia, by affecting the mix of inflation volatility and debt-deflation risk in the economy.

C0269: **The interplay between financial conditions and monetary policy shocks**
Presenter: **Luca Benzoni**, Federal Reserve Bank of Chicago, United States
Co-authors: Marco Bassetto, Trevor Serrao
An empirical study is conducted concerning the interplay between monetary policy and financial conditions shocks. We find that such shocks have a significant and similar impact on the real economy, though with different degrees of persistence, and the systematic fed funds rate response to a financial shock contributes to bringing the economy back towards its trend. However, a binding zero lower bound on policy rates can prevent policy-makers from leaning against the wind, with a significant cost in terms of output and investment. We illustrate these conclusions in a retrospective analysis of the U.S. economy over the past 20 years, in which we decompose the realization of economic variables into the contributions of financial, monetary policy, and other shocks.

CO192 Room MAL 538 REAL-TIME DATA ANALYSIS Chair: Simon van Norden

C0978: **The effect of recessions on potential output estimates: Size, timing, and determinants**
Presenter: **Jonas Dovern**, Heidelberg University, Germany
Co-authors: Christopher Zuber
The aim is to analyze when and how much potential output estimates are revised following recessions. We find that they are revised downward, revisions are larger after supply-driven recessions, the size of revisions are in line with results obtained by statistical filters, average revisions are equal in size to those obtained in a simulation with no hysteresis, and the recession depth and pre-recession values of the primary balance, the current account, and credit volumes are significant predictors of post-recession revisions. Our results call for improved methods for estimating potential output and suggest that average recessions do not generate substantial hysteresis effects.

C0939: **Early warning systems for currency crises with real-time data**
Presenter: **Jan Jacobs**, University of Groningen, Netherlands
Co-authors: Tjeerd Boonman, Gerard Kuper, Alberto Romero
The performance of early warning systems for currency crises in real-time is investigated by using forecasts of indicators that are available at the moment predictions are to be made. We focus on eight Latin American and Central and Eastern European countries, distinguishing an estimation period 1990-2009 and a prediction period 2010-2014. We apply two varieties of early warning systems: the signal approach and the logit models. For both methods we find that using early estimates in the predictions worsens the ability of early warning systems to signal crises compared to the most recently available information.

C1191: **Advances in nowcasting economic activity**
Presenter: **Thomas Drechsel**, LSE, United Kingdom
Co-authors: Ivan Petrella, Juan Antolin Diaz
The nowcasting performance of a Dynamic Factor Model which extends the existing framework along various dimensions is evaluated. First, building on our prior work, we incorporate low-frequency movements to the growth rate and the volatility of the variables. Second, we consider the impact of relaxing the assumption of homogeneous dynamics by allowing for lags of the factor to enter the measurement equation. Third, we endogenously model seasonal fluctuations, which can be particularly useful whenever there is suspicion that residual seasonality is present in the data, and outliers, by allowing for fat tailed distributions in the idiosyncratic disturbances. Using a fully real-time database covering the period 2000-2017 for the G7 economies, we conduct a comprehensive forecasting evaluation exercise, with particular attention to probabilistic (density) forecasts, and assess the importance of these features for the real time tracking of economic activity.

CO280 Room MAL 539 NEW CHALLENGES FOR SEASONAL AND CALENDAR ADJUSTMENT Chair: Gian Luigi Mazzi

C0844: **Impact of atypical weather of seasonal adjustment**
Presenter: **Susie Fortier**, Statistics Canada, Canada
Co-authors: Steve Matthews, Zdenek Patak

One of the challenges of seasonal adjustment is unseasonal events such as weather extremes. It is conceivable that long stretches of unusually low or high temperatures, prolonged drought conditions, extended rainfall, to name a few, may impact the behaviour of both individuals and businesses. Seasonal adjustment removes repeating, equally spaced patterns from the corresponding time series data, along with movement due to various holidays and other predefined regressors, but it does not remove the impact that atypical weather may have. To fill the void, many national statistical agencies have been studying the use of weather related indicators in their time series analyses, including seasonal adjustments. We discuss alternatives for incorporating the data into seasonal adjustment and ad hoc analysis.

C0879: **Two-step reconciliation of time series**
Presenter: **Enrico Infante**, European Commission - Eurostat, Luxembourg
Co-authors: Germana Scepi

Two-step reconciliation methods solve the temporal constraint in the first step, while in the second step the contemporaneous constraint is satisfied without altering the temporal constraint. Previous methodology applies the Denton benchmarking technique in the first step. Based on an alternative two-step procedure for the reconciliation of systems of time series, an algorithm is proposed which allows us to choose one of the two different solutions for the second step, and introduces the possibility of using well-known established techniques in the first step. Furthermore, a way of dealing with the reconciliation of hierarchical systems of time series is presented. An innovative test for detecting common seasonal patterns in time series is also presented. Such test could be used for deciding at which level to seasonally adjust an aggregated time series before applying reconciliation. Moreover, together with a simulation study, several aspects of the validation of a reconciliation technique are shown, including a new methodology for detecting whether the outliers at the end of series are consistent. Two real examples using the European industrial production index and the euro area quarterly sector accounts data will also be presented.

C0926: **Seasonal adjustment of daily data with JDemetra+: New results**
Presenter: **Dominique Ladiray**, INSEE, France
Co-authors: Tommaso Proietti, Gian Luigi Mazzi

The progresses in information technology have fostered the availability of daily and weekly time series. The seasonal adjustment of high-frequency time series poses several challenges. First of all the seasonal period of the annual cycle is neither constant nor an integral. Secondly, in order to accommodate complex seasonal patterns many individual effects might be required. Thirdly, the need for robust methods is reinforced by the fact that the effects of outlying observations is not smoothed by temporal aggregation and that they are relatively more frequent. Moving average methods (X-13ARIMA-SEATS) and ARIMA-based methods (TRAMO-SEATS) are recommended by Eurostat and are implemented in JDemetra+, the European software for seasonal adjustment. These methods only deal with monthly and quarterly series; they estimate first the outliers and calendar effects using a Reg-ARIMA model and then decompose the residual of the model into trend-cycle, seasonality and irregular component.We will present implementations of a Tramo-Seats algorithm and a X-12 algorithm to seasonally adjust high frequency data with multiple and non-integer seasonalities.

CO202 Room MAL 540 BAYESIAN NONLINEAR ECONOMETRICS Chair: Roberto Casarin

C1225: **Bayesian Markov switching tensor regression for time-varying networks**
Presenter: **Roberto Casarin**, University Ca Foscari of Venice, Italy
Co-authors: Monica Billio, Matteo Iacopini

A Bayesian Markov Switching regression is proposed for modelling dynamic multilayer networks. We apply a suitable low-rank decomposition of the tensor of coefficients for parsimony and avoiding the over-fitting. The time-varying parameters are driven by a hidden Markov chain whose states are identified by means of constraints imposed on the regime specific parameters which determine the number of observed edges. In addition, we jointly model a vector of observables. We exploit the Polya-Gamma data augmentation scheme for logit models in order to provide an efficient Gibbs sampler for posterior inference. We show the effectiveness of the sampler on simulated and real datasets.

C1217: **Bayesian dynamic tensor regression**
Presenter: **Matteo Iacopini**, Ca Foscari University of Venice, Italy
Co-authors: Roberto Casarin, Monica Billio

A new dynamic linear regression model is proposed for tensor variate response and covariates that encompasses univariate, multivariate (i.e. SUR, VAR, panel VAR) and matrix regression models as special cases. For dealing with the over-parametrization and over-fitting issues due to the curse of dimensionality, we exploit a suitable parametrization which enables to achieve both parameter parsimony and to incorporate sparsity effects. Inference is carried out in the Bayesian framework combined with Monte Carlo Markov Chain (MCMC). We show the eficiency of the MCMC procedure on simulated data, then we apply our methodology on macroeconomic and financial real datasets.

C1228: **A discrete-time stochastic volatility framework for pricing options with realized measures**
Presenter: **Giulia Livieri**, Scuola Normale Superiore, Italy
Co-authors: Roberto Casarin, Giacomo Bormetti, Fulvio Corsi

Motivated by the fact that realized measures of volatility are affected by measurement errors, we introduce a new family of discrete-time stochastic volatility models having two measurement equations relating both the observed returns and realized measures to the latent conditional variance. A fully analytical option pricing framework is developed for this new discrete-time Stochastic Volatility class of models. In addition, we provide analytical filtering and smoothing recursions for the basic version of the model, and an effective MCMC algorithm for its richer variants. The empirical analysis shows the efficacy of the filtering and smoothing of realized measures in inating the latent volatility persistence - the crucial parameter for the effective pricing of Standard and Poor's 500 Index options.

CO647 Room MAL 541 FINANCIAL ECONOMETRICS THEORY, INSURANCE, AND RISK MANAGEMENT Chair: Debbie Dupuis

C0945: **Valuation of guaranteed unitized participating life insurance under MEGB2 distribution**

Presenter: **Haitao Zheng**, Beihang University, China

Crisis events have significantly changed the view that extreme events in financial markets have negligible probability. Especially in the life insurance market, the price of guaranteed participating life insurance contract will be affected by a change in asset volatility which leads to the fluctuations in embedded option value. Considering the correlation of different asset prices, the MEGB2 (Multivariate exponential generalized beta of the second kind) distribution is proposed to price guaranteed participating life insurance contract which can effectively describe the dependence structure of assets under some extreme risks. Assuming the return rates of two different assets follow the MEGB2 distribution, a multi-factor fair valuation pricing model of insurance contract is split into four components: the basic contract, the annual dividend option, the terminal dividend option and the surrender option. The aim is to study the effect of death rate, minimum guaranteed interest rate, annual dividend ratio, terminal dividend ratio and surrender on the embedded option values and calculates the single premium of the insurance contract under different influence factors. The Least-Squares Monte Carlo simulation method is used to simulate the pricing model. Finally, a comparison is made for the sensitivity of the pricing parameters under the MEGB2 distribution and Multivariate Normal distribution.

C0885: **Joint extrapolation forecasting of supply and use tables based on matrix transformation techniques**

Presenter: **Cheng Wang**, Beihang University, China

Co-authors: Huiwen Wang, Qi Fang

Plenty of methods have been proposed to update Supply and Use tables (SUTs). They rely on the assumption that the economic structure will not change significantly during the interpolation period. However, this assumption may not always hold, particularly for countries experiencing rapid development. Besides, they require the availability of the use and supply totals by products, which leads to that they cannot extrapolation forecasting SUTs. The aim is to combine forecasting with a matrix transformation technique (MTT) to provide a new perspective on joint forecasting SUTs. Under the assumption that changes in the trend of an economic structure are statistically significant, the method extrapolates SUTs by combining time series models with MTT and proceeds without economic data for the target years. A simulation study and an empirical analysis are conducted to compare the forecasting performance of the MTT to the Generalized RAS (GRAS)method. The results show that when we know the use and supply totals by products, it can be proved that the comprehensive performance of the MTT is as goog as the performance of the GRAS method, which shows the effectiveness of MTT. Thus MTT can demonstrate its merits to extrapolation forecasting SUTs when GRAS method cannot be used.

C0250: **Scenario response distributions**

Presenter: **Fulvio Pegoraro**, European Central Bank and CREST, France

Co-authors: Caroline Jardet, Alain Monfort

A statistical methodology generalizing the standard Impulse Response Function (IRF) theory in two directions is proposed. First, we consider the case where the new information may be not only about the present value of the multivariate process of interest, but also about future values. Second, we show how to compute not only the average responses of the variables but the whole distribution of these responses and, therefore, several features like variances, quantiles or prediction intervals. Our methodology, called Scenario Response Distribution (SRD) methodology, can be used not only in the context of standard Gaussian VAR models, but also in the context of nonlinear models, in particular censored or regime switching models. An empirical exercise illustrating the SRD methodology will be provided.

Authors Index

