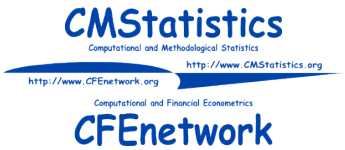
PROGRAMME AND ABSTRACTS

2nd International Conference on
Econometrics and Statistics (EcoSta 2018)

http://cmstatistics.org/EcoSta2018

City University of Hong Kong
19 – 21 June 2018



ISBN: 978-9963-2227-3-5

Co-chairs:
Igor Pruenster, Alan Wan, Ping-Shou Zhong.

EcoSta Editors:
Ana Colubi, Erricos J. Kontoghiorghes, Manfred Deistler.

Scientific Programme Committee:
Tomohiro Ando, Jennifer Chan, Cathy W.S. Chen, Hao Chen, Ming Yen Cheng, Jeng-Min Chiou, Terence Chong, Fabrizio Durante, Yingying Fan, Richard Gerlach, Michele Guindani, Marc Hallin, Alain Hecq, Daniel Henderson, Robert Kohn, Sangyeol Lee, Degui Li, Wai-Keung Li, Yingying Li, Hua Liang, Tsung-I Lin, Shiqing Ling, Alessandra Luati, Hiroki Masuda, Geoffrey McLachlan, Samuel Mueller, Yasuhiro Omori, Marc Paolella, Sandra Paterlini, Heng Peng, Artem Prokhorov, Jeroen Rombouts, Matteo Ruggiero, Mike K.P. So, Xinyuan Song, John Stufken, Botond Szabo, Minh-Ngoc Tran, Andrey Vasnev, Judy Huixia Wang, Yong Wang, Yichao Wu and Jeff Yao.

Local Organizing Committee:
Guanhao Feng, Daniel Preve, Geoffrey Tso, Inez Zwetsloot, Catherine Liu, Zhen Pang.

Dear Colleagues,

It is a great pleasure to welcome you to the 2nd International Conference on Econometrics and Statistics (EcoSta 2018). The conference is co-organized by the working group on Computational and Methodological Statistics (CMStatistics), the network of Computational and Financial Econometrics (CFEnetwork), the journal Econometrics and Statistics (EcoSta) and the Department of Management Sciences of the City University of Hong Kong (CityU).

Following the success of the first edition, the aim is for the conference to become a leading meeting in econometrics, statistics and their applications.

The EcoSta 2018 consists of about 140 sessions, three keynote talks, three invited sessions, and 550 presentations. There are over 600 participants. These numbers confirm the support of the involved research communities to this important initiative. It is indeed promising that the EcoSta conference will become a successful medium for the dissemination of high quality research in Econometrics and Statistics, and facilitate networking.

The Co-chairs acknowledge the collective effort of the scientific program committee, session organizers, and local organizing committee, which has produced a programme that spans all the areas of econometrics and statistics. The CityU provides excellent facilities and a fantastic environment conveniently located in the center of Hong Kong. The local host, volunteers, and sponsoring universities have substantially contributed through their effort to the successful organization of the conference. We thank them all for their support. Particularly we express our sincere appreciation to the host and main sponsor, the Department of Management Sciences of the CityU.

It is hoped that the quality of both the scientific programme and the CityU will provide the participants with a productive, stimulating conference, and an enjoyable stay in Hong Kong.

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2018 conference. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta and CSDA, and to join the networks.

Finally, we are happy to announce that the 3rd International Conference on Econometrics and Statistics (EcoSta 2019) will take place at the National Chung Hsing University, Taiwan from Tuesday the 25th to Thursday the 27th of June 2019. You are invited to participate actively in these events. Tutorials will take place on Friday the 28th of June 2019.

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

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

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

Specialized teams

Currently the ERCIM WG has over 1650 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

2018-06-19	2018-06-20	2018-06-21
Opening, 08:45 - 09:00	**F** EcoSta2018 08:30 - 09:50	**J** EcoSta2018 08:30 - 10:10
A - Keynote EcoSta2018 09:00 - 09:50	**Coffee Break** 09:50 - 10:25	**Coffee Break** 10:10 - 10:40
Coffee Break 09:50 - 10:25	**G** EcoSta2018 10:20 - 12:25	**K** EcoSta2018 10:40 - 12:20
B EcoSta2018 10:25 - 12:30	**Lunch Break** 12:25 - 14:00	**Lunch Break** 12:20 - 13:50
Lunch Break 12:30 - 14:00	**H** EcoSta2018 14:00 - 15:40	**L** EcoSta2018 13:50 - 15:30
C EcoSta2018 14:00 - 15:40	**Coffee Break** 15:40 - 16:10	**Coffee Break** 15:30 - 16:00
Coffee Break 15:40 - 16:10	**I** EcoSta2018 16:10 - 17:50	**M** EcoSta2018 16:00 - 17:15
D EcoSta2018 16:10 - 17:25	**Conference Dinner** 19:30 - 22:15	**N - Keynote** EcoSta2018 17:25 - 18:15
E - Keynote EcoSta2018 17:40 - 18:30		**Closing**, 18:15 - 18:30
Welcome Reception 18:35 - 20:00		

REGISTRATION AND SOCIAL EVENTS

- *Registration*. The registration will be open on Monday the 18th of June 2018, 13:30 - 18:00, and during the days of the conference 8:00-18:00 at the University Concourse of the 4th floor of the Yeung Kin Man Academic Building **(see maps and indications on pages VIII-X)**. The conference badges have a QR code with the registration information of the participants. For this reason, it is mandatory to always bring the conference badge.
- *The coffee breaks* will take place at the University Concourse of the 4th floor of the Yeung Kin Man Academic Building **(see maps and indications on pages VIII-X)**. You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Tuesday the 19th of June 2018, from 18:35 - 20:00*. The Welcome Reception is open to the conference participants. It will take place at the space in front of the Wong Cheung Lo Hui Yuet Hall, 5th Floor of the Lau Ming Wai Academic Building **(see maps and indications on pages VIII-X)**. Conference registrants must bring their conference badge in order to attend the reception. Preregistration is required due to health and safety regulations, and limited capacity of the venue. Entrance to the reception venue will be strictly limited to those who have registered.
- *Conference Dinner, Wednesday the 20th of June 2018, from 19:30 to 22:15*. The conference dinner is optional and registration is required. It will take place at the Colour Crystal Restaurant, 2/F, Harbour Crystal Centre, Tsim Sha Tsui East **(see maps at page XI)**. Conference registrants must bring their conference badge in order to attend the conference dinner.
- *Lunches*. Participants can buy lunch at restaurants and cafes at the CityU or at the Mall (Festival Walk) by the conference venue. A list of suggested places is at the conference website (http://cmstatistics.org/EcoSta2018/lunches.php).

TUTORIAL

A tutorial on Functional Data Analysis will be given by Prof. Jane-Ling Wang, University of California, Davis. The tutorial will take place on Monday the 18th of June 2018 from 14:00 to 18:30 at Room LT-14, 4th floor of the Yeung Kin Man Academic Building **(see maps and indications on pages VIII-X)**. Pre-registration is required.

SPECIAL MEETINGS by invitation to group members

- The *Econometrics and Statistics (EcoSta) Editorial Board* meeting will take place on Monday the 18th of June 2018, 17:15-18:15, Room G4701 **(see maps and indications on pages VIII-X)**. The meeting is by invitation only.
- The *Econometrics and Statistics (EcoSta) Editorial Board* dinner will take place on Monday the 18th of June 2018, 19:00-21:00. The meeting point will be the registration desk 18:15-18:45 **(see maps and indications on pages VIII-X)**. The dinner is by invitation only.

GENERAL INFORMATION

Addresses of venues (see maps and indications on pages VIII-X)

- The registration, tutorial, coffee breaks and parallel sessions will take place at the 4th floor of the Yeung Kin Man Academic Building (AC1) of the CityU, Tat Chee Avenue, Kowloon, Hong Kong.
- The 1st and 2nd keynote talks and the Welcome Reception will take place at the 5th Floor of the Lau Ming Wai Academic Building (AC3) of the CityU, Tat Chee Avenue, Kowloon, Hong Kong. The 3rd keynote talk will take place at the 4th floor of the Yeung Kin Man Academic Building (AC1) of the CityU, Tat Chee Avenue, Kowloon, Hong Kong.
- The Conference Dinner will take place at the Colour Crystal Restaurant, 2nd floor of the Harbour Crystal Centre, Tsim Sha Tsui East, Hong Kong.

Lecture rooms (see maps and indications on pages VIII-X)

The opening and keynote talks 1 and 2 will take place at the Wong Cheung Lo Hui Yuet Hall, 5th Floor of the Lau Ming Wai Academic Building (AC3). The kyenote talk 3 will take place at room LT-1 of the Yeung Kin Man Academic Building (AC1). The paper and poster presentations will take place at the 4th floor of the Yeung Kin Man Academic Building (AC1) of the CityU. We advise that you visit the venue in advance.

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 point), 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) 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 registration desk area, 4th floor of the Yeung Kin Man Academic Building (AC1). 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.

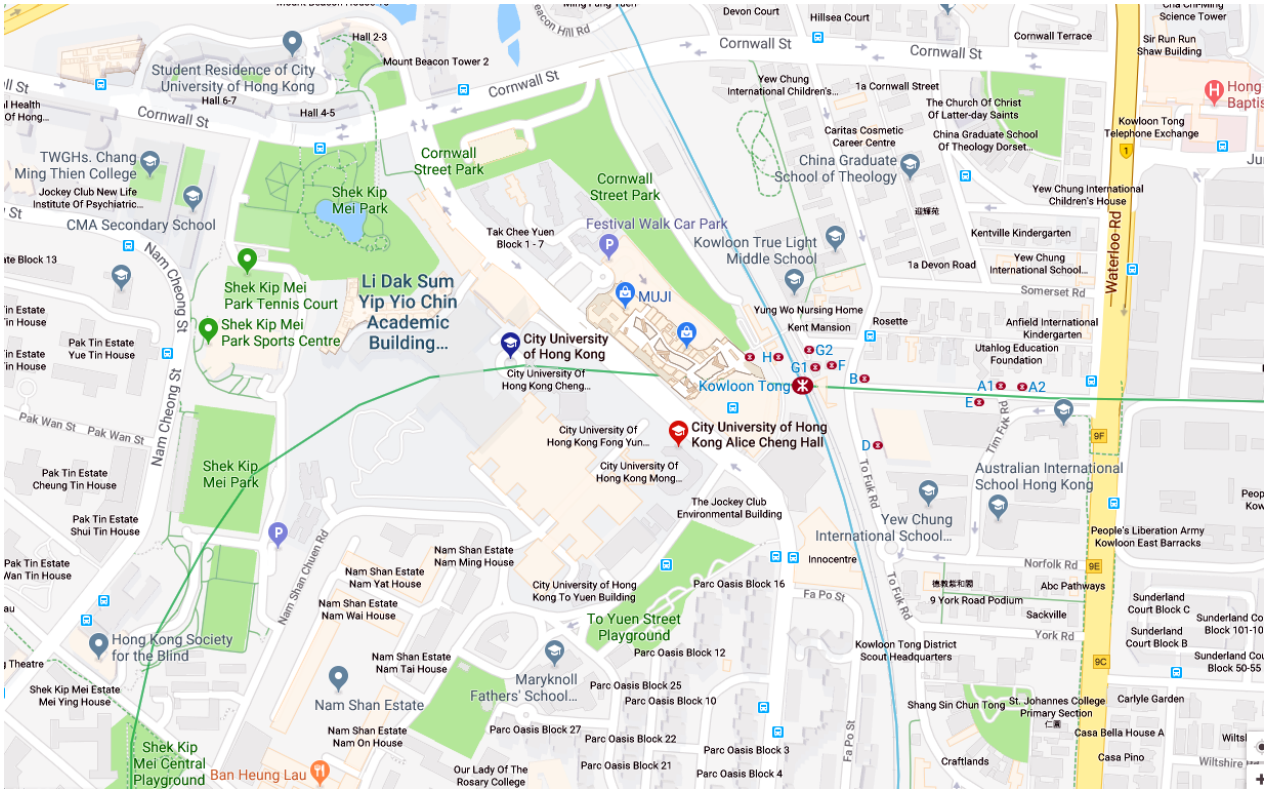
Internet connection

The information for the wireless internet connection will be displayed by the registration desk.

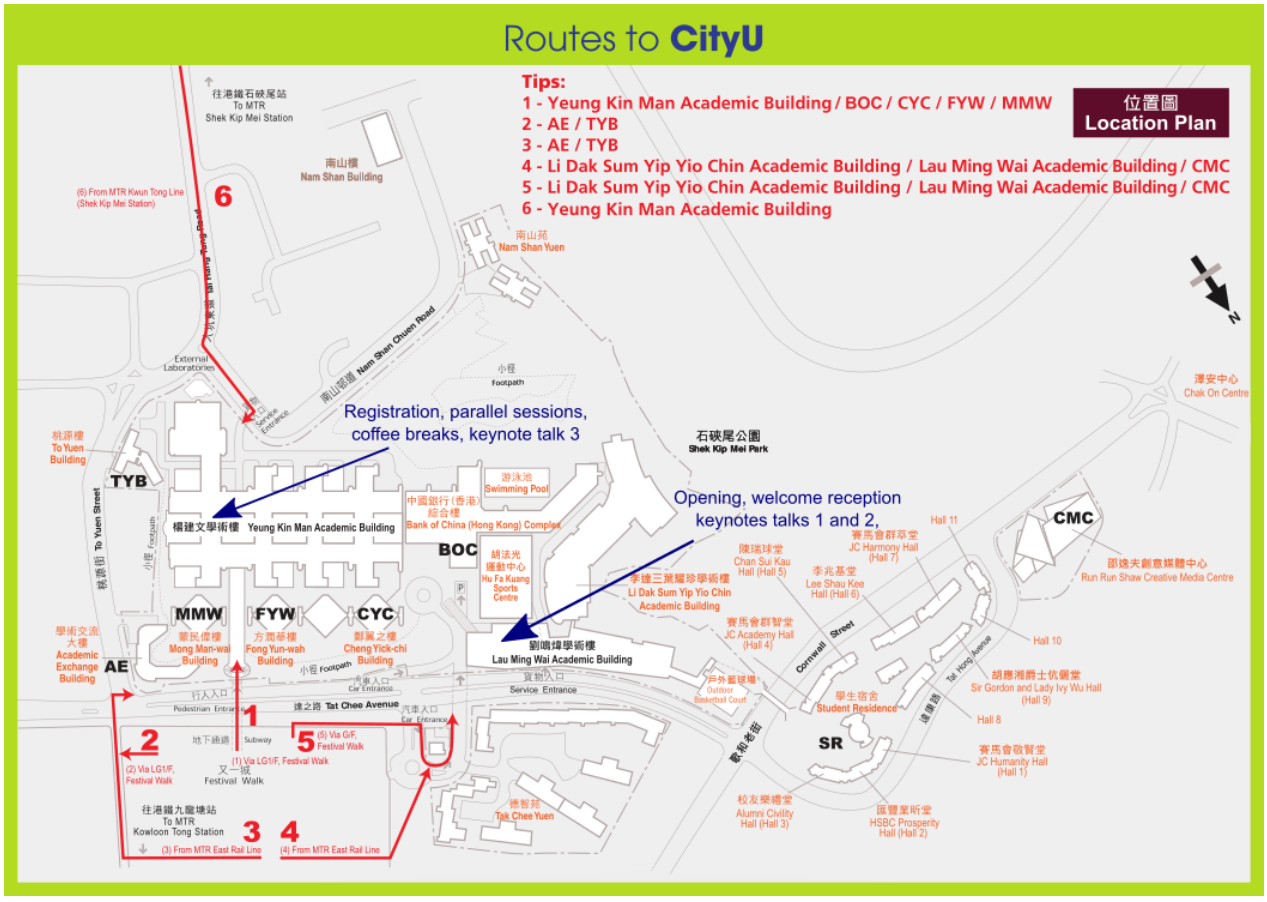
Exhibitors

Elsevier.

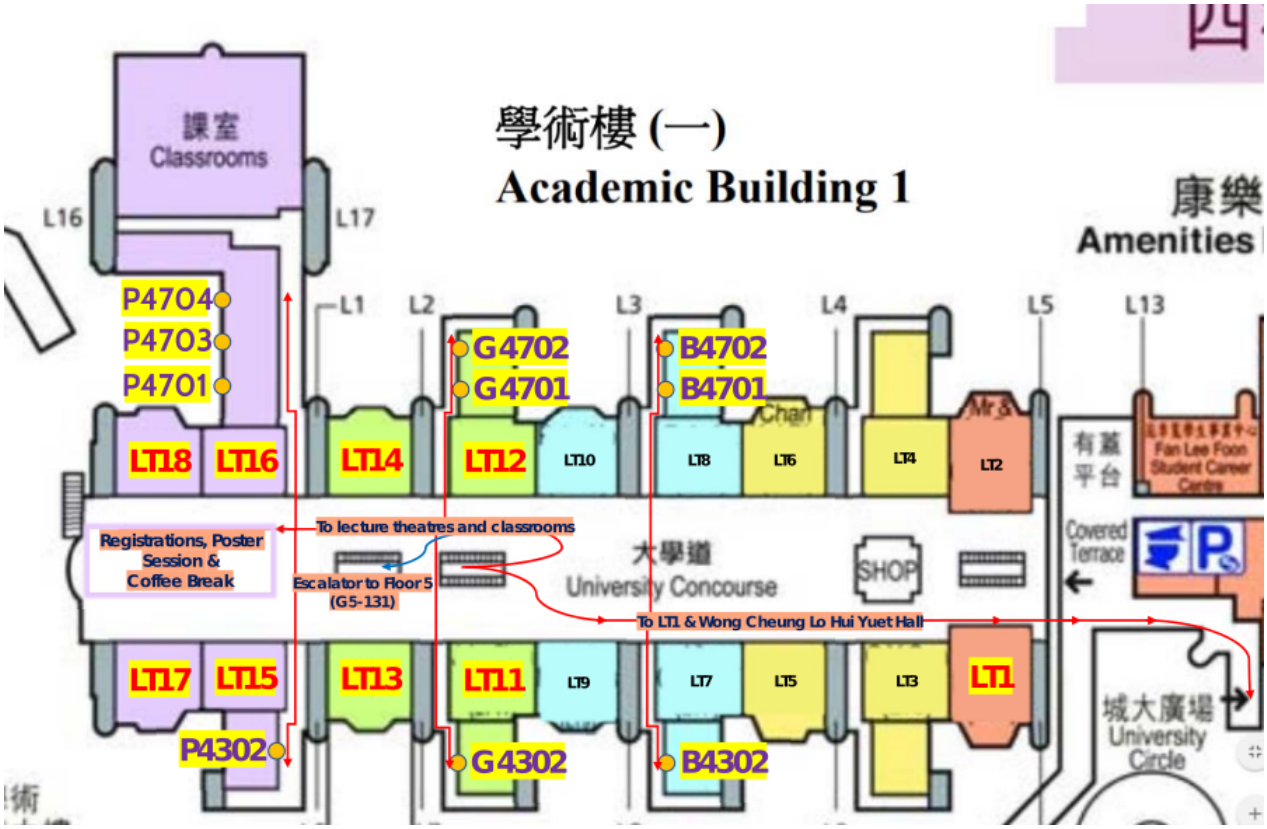
Google map of the venue and nearby area



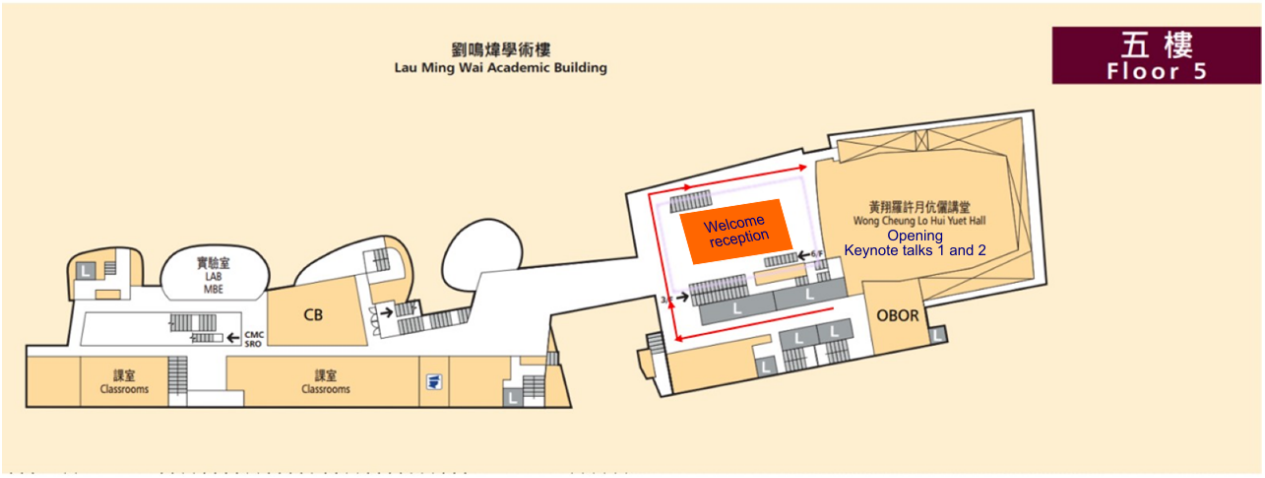
Routes from the Metro Station (locally known as Mass Transit Railway -MTR-) Kowloon Tong



Venue: 4th floor of the Yeung Kin Man Academic Building (AC1)



Venue: 5th floor of the Lau Ming Wai Academic Building (AC3)



Lecture Theatres, 4th floor Yeung Kin Man Academic Building & Wong Chueng Lo Hui Yuet Hall

<table>
<tr><th rowspan="2">Venue</th><th rowspan="2">Capacity</th><th colspan="4">Steps to arrive at different venues</th></tr>
<tr><th>① Pass through the Pedestrian Subway</th><th>② Enter the building (AC1)</th><th>③ In University Concourse (4/F, AC1)</th><th>④ Arrive at the venue</th></tr>
<tr><td>LT-1</td><td>300</td><td rowspan="9">When you get off the MTR, look for Festival Walk exit.
↓
In Festival Walk, on Level LG1, there is a Pedestrian Subway which will lead you to CityU campus.
↓
Go straight after walking through the Pedestrian Subway, walk through the red doors to enter the Yeung Kin Man Academic Building.</td><td rowspan="9">Go straight ahead past the Bookshop and you will see escalators on your right.
↓
Go up one level to the University Concourse.</td><td rowspan="8">You will see all the lecture theatres on both sides of the Concourse.</td><td rowspan="8">4/F University Concourse, Yeung Kin Man Academic Bld. (4/F, AC1)</td></tr>
<tr><td>LT-11</td><td>120</td></tr>
<tr><td>LT-12</td><td>120</td></tr>
<tr><td>LT-13</td><td>144</td></tr>
<tr><td>LT-14</td><td>140</td></tr>
<tr><td>LT-15</td><td>120</td></tr>
<tr><td>LT-16</td><td>120</td></tr>
<tr><td>LT-17</td><td>200</td></tr>
<tr><td>LT-18</td><td>200</td></tr>
<tr><td>Registration, poster session and coffee breaks</td><td>--</td><td>You will see the registration area in front of you.</td><td>4/F University Concourse, Yeung Kin Man Academic Bld. (4/F, AC1)</td></tr>
<tr><td>Wong Cheung Lo Hui Yuet Hall</td><td>600</td><td>Go along the University Concourse and exit the building, you will find the University Circle on your right hand side.</td><td>Go along the covered walkway and follow the directional signs which will lead you to 5/F, Lau Ming Wai Academic Building.</td></tr>
</table>

Classrooms, 4th floor Yeung Kin Man Academic Building

<table>
<tr><th rowspan="2">Venue</th><th rowspan="2">Capacity</th><th colspan="4">Steps to arrive at different venues</th></tr>
<tr><th>① Pass through the Pedestrian Subway</th><th>② Enter the building (AC1)</th><th>③ In University Concourse (4/F, AC1)</th><th>④ Arrive at the venue</th></tr>
<tr><td>P4302</td><td>80</td><td rowspan="13">When you get off the MTR, look for Festival Walk exit.
↓
In Festival Walk, on Level LG1, there is a Pedestrian Subway which will lead you to CityU campus.
↓
Go straight after walking through the Pedestrian Subway, walk through the red doors to enter the Yeung Kin Man Academic Building.</td><td rowspan="13">Go straight ahead past the Bookshop and you will see escalators on your right.
↓
Go up one level to the University Concourse.</td><td rowspan="13">You will see zones divided into different colours.</td><td>Walk to the LEFT hand side of the PURPLE zone.</td></tr>
<tr><td>P4701</td><td>80</td><td rowspan="3">Walk to the RIGHT hand side of the PURPLE zone.</td></tr>
<tr><td>P4703</td><td>80</td></tr>
<tr><td>P4704</td><td>68</td></tr>
<tr><td>B4302</td><td>80</td><td>Walk to the LEFT hand side of the BLUE zone.</td></tr>
<tr><td>B4701</td><td>43</td><td rowspan="2">Walk to the RIGHT hand side of the BLUE zone.</td></tr>
<tr><td>B4702</td><td>40</td></tr>
<tr><td>G4302</td><td>80</td><td>Walk to the LEFT hand side of the GREEN zone.</td></tr>
<tr><td>G4701</td><td>43</td><td rowspan="2">Walk to the RIGHT hand side of the GREEN zone.</td></tr>
<tr><td>G4702</td><td>43</td></tr>
</table>

Google map from CityU to the Conference Dinner Restaurant

Google map from Hung Hom MTR station to the Conference Dinner Restaurant

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 (CFEnetwork) and Computational and Methodological Statistics (CMStatistics). 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. Well-founded applied econometric studies that demonstrate the practicality of new procedures and models are of interest as well. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.

Part B: Statistics. Papers providing important original contributions to methodological statistics inspired by applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering. In general, the interaction of mathematical methods and numerical implementations for the analysis of large and/or complex datasets arising in areas such as medicine, epidemiology, biology, psychology, climatology and communication is considered. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them, as complementary material.

The journal consists, preponderantly, of original research. Occasionally, reviews and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

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

Papers 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:

- (Part A: Econometrics) Annals of Computational and Financial Econometrics
- (Part A: Econometrics) Special Issue on Theoretical Econometrics.
- (Part A: Econometrics) Special Issue on Computational Econometrics.
- (Part B: Statistics) Special Issue on Copulas.
- (Part B: Statistics) Special Issue on Neuroimaging.

The deadline for paper submissions is the 15th July 2018. 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 the journal Computational Statistics & Data Analysis (CSDA). Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/csda. Any questions may be directed via email to: csda@dcs.bbk.ac.uk.

Contents

Tuesday 19.06.2018 09:00 - 09:50 Room: Wong Cheung Lo Hui Yuet Hall Chair: Ana Colubi Keynote talk 1

Functional mixed effects models for longitudinal functional responses

Speaker: **Jane-Ling Wang, University of California Davis, United States** Hongtu Zhu, Kehui Chen, Xinchao Luo, Ying Yuan

Longitudinal functional data consist of functional data collected at multiple time points for which the observational times may vary by subject. They differ from traditional longitudinal data in that the observation at each time point is a function rather than a scalar. The aim is to extend the traditional linear mixed-effects model for longitudinal data to longitudinal functional data. We study a class of functional mixed effects models(FMEM), which includes fixed effects that characterize the association between longitudinal functional responses and covariates of interest and random effects that capture the intricate correlation structure of longitudinal functional responses. We propose local linear estimates for the fixed-effect coefficient functions and establish their asymptotic properties. We also develop a simultaneous confidence band for each fixed-effect coefficient function and a global test for linear hypotheses of these coefficient functions. The numerical performance of the proposed methods is examined through an extensive simulation study and an application to white-matter fiber data from a national database for autism research.

Tuesday 19.06.2018 17:40 - 18:30 Room: Wong Cheung Lo Hui Yuet Hall Chair: Alan Wan Keynote talk 2

Selection of an optimal rolling window in time-varying predictive regression

Speaker: **Yongmiao Hong, Cornell University, United States** Yuying Sun, Shouyang Wang

Since the underlying economic structure is likely to be affected by changes in preferences, technologies, policies, crises, etc., data in the previous time period may be irrelevant to the present data-generation process. Thus, econometric forecasts are often based on rolling estimation. However, it is far from clear how to choose an optimal sample to estimate a predictive model. We propose a novel approach to selecting the optimal window in a predictive linear regression model with time-varying parameters, by minimizing suitable criteria based on forecast errors, including unconditional/conditional/global mean square forecast errors. A practically feasible cross-validation procedure is developed to choose an optimal window, which is asymptotically equivalent to the infeasible optimal window based on unconditional mean square forecast errors. Simulation studies are conducted to evaluate the accuracy of forecasts using our methods under various types of structural changes. Empirical applications, to forecasting the US GDP growth rates, inflation rates and stock returns, highlight the merits of the proposed methods relative to other popular methods available in the literature.

Thursday 21.06.2018 17:25 - 18:15 Room: LT-1 Chair: Mike So Keynote talk 3

On choosing mixture components via non-local priors

Speaker: **Mark Steel, University of Warwick, United Kingdom** David Rossell, Jairo Fuquene

Choosing the number of mixture components remains a central but elusive challenge. Traditional model selection criteria can be either overly liberal or conservative when enforcing parsimony. They may also result in poorly separated components of limited practical use. Non-local priors (NLPs) are a family of distributions that encourage parsimony by enforcing a separation between the models under consideration. We formalize NLPs in the context of mixtures and show how they lead to well-separated components that are interpretable as distinct subpopulations. We suggest default prior settings, give a theoretical characterization of the sparsity induced by NLPs, derive tractable expressions and propose simple algorithms to obtain the integrated likelihood and parameter estimates. The framework is generic and we fully develop multivariate Normal, Binomial and product Binomial mixtures based on a family of exchangeable moment priors. Our results show a serious lack of sensitivity of the Bayesian information criterion (BIC) and insufficient parsimony of the AIC and a local prior counterpart to our formulation. The singular BIC behaved like the BIC in some examples and the AIC in others. We also offer comparisons to overfitted and repulsive overfitted mixtures, the performance of which depended on the choice of prior parameters. The number of components inferred under NLPs was closer to the true number (when known)and remained robust to prior settings.

Tuesday 19.06.2018 08:45 - 09:00 Room: Wong Cheung Lo Hui Yuet Hall Chair: Alan Wan Opening

Opening speach

Speaker: **Houmin Yan, City University of Hong Kong, China**

Prof. Houmin YAN, Dean, School of Business, City University of Hong Kong, welcomes all participants to EcoSta 2018. In his opening speech, Prof. YAN will give an overview of the university's Business School, its history, mission and recent achievements.

Tuesday 19.06.2018 10:25 - 12:30 Parallel Session B – EcoSta2018

EI002 Room LT-17 RECENT ADVANCES IN NONPARAMETRIC STATISTICS Chair: Alan Wan

E0157: **Nonparametric multi-dimensional fixed effect panel data models**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Juan Manuel Rodriguez-Poo, Alexandra Soberon
Multi-dimensional panel data sets are becoming increasing popular to identify marginal effects in empirical research. Fixed effects estimators are typically employed in order to deal with potential correlation between unobserved effects and regressors. Nonparametric estimators for one-way fixed effects models exist, but are cumbersome to employ in practice as they typically require iteration, marginal integration or profile estimation. We develop a nonparametric estimator for the gradient of the conditional mean that works for any dimension fixed effect model and has a closed-form solution. The asymptotic properties of our estimator are given and the finite sample properties are shown via simulations, as well as via an empirical application which further extends our estimator to the partially linear setting.

E0526: **DNN: A two-scale distributional tale of causal inference**
Presenter: **Jinchi Lv**, University of Southern California, United States
Co-authors: Yingying Fan, Jingbo Wang
The problem of heterogeneous treatment effect estimation and inference in nonparametric regression models is considered. To reduce the bias in the k-nearest neighbors estimation method, we exploit the idea of subsampling. Then, by comparing and contrasting the k-nearest neighbors estimators with different subsampling scales, we are able to successfully achieve desired bias reduction. Under some mild regularity conditions, the resulting new DNN estimator is proved to be asymptotically unbiased and have asymptotically normal distribution. The method and theoretical results are supported by several simulation examples. The approach is also applied to a child's birth weight data set to study the heterogeneous treatment effect of smoking.

E0796: **Estimating shape constrained functions using Gaussian processes**
Presenter: **Xiaojing Wang**, University of Connecticut, United States
Co-authors: Jim Berger
In many applications, economic theory often provides shape restrictions on functions of interest, such as the option pricing function must be monotonic and convex, utility function associated with rational preference should be monotone and so on. Whenever the economists want to model such economic relationships, they not only have to consider economic theory, but also have to take account of flexibility of functional forms. This motivates us to consider nonparametric modeling of functional relationships between economic variables under shape restrictions. Gaussian processes are a popular tool for nonparametric function estimation because of their flexibility and the fact that much of the ensuing computation is parametric Gaussian computation. The shape constraints can be then incorporated through the use of derivative processes, which are joint Gaussian processes with the original process, as long as the conditions of mean square differentiability hold. The possibilities and challenges of introducing shape constraints through Gaussian processes models are explored, and illustrated through simulations and real data examples. Computation is carried out through a Gibbs sampling scheme.

EO308 Room B4302 ALTERNATIVE RISK PREMIA Chair: Serge Darolles

E0565: **Alternative risk premia: Benchmarking and performance evaluation**
Presenter: **Guillaume Monarcha**, Orion Financial Partners, France
Through the analysis of more than 400 investable ARPs from 10 investment banks, we show that ARPs investing is more challenging than simply premia harvesting. First, if 2/3 of the investable ARPs are academic-based, more than 1/3 rely on proprietary quantitative trading strategies, and therefore do not rely on academic factors. Second, a priori similar ARPs may exhibit significant divergences in their distributional properties, as well as in their correlation structure with academic ARPs, other investable ARPs, and traditional asset classes. These divergences are linked to specific features in their construction methodologies: factor definition, nature of the short leg of the strategy, allocation methodology, selection criteria, and specific parametrization. Third, 80% of the investable ARPs are effectively investable since late 2015, implying that their historical track records are mainly based on backtests, and therefore potentially encompass significant backtesting bias. To assess these issues, we propose a three-step analysis framework dedicated to the performance evaluation of ARPs. First, we propose to identify the main ARP styles through clustering. Second, we build ARP benchmarks based on the extraction and the identification of the common latent factors that drive each ARP style. Third, we propose a dynamic performance evaluation model, that accounts for potential backtesting bias.

E0493: **Deep learning alpha**
Presenter: **Guanhao Feng**, City University of Hong Kong, Hong Kong
Co-authors: Nicholas Polson, Jianeng Xu
The goal is to push the classical long-short portfolio asset pricing framework to the future of artificial intelligence. Sorting securities on firm characteristics and constructing long-short portfolios is a tradition to both the asset pricing academia and hedge fund industry. Sorting is a nonlinear activation that can be built within a deep learning architecture and works for the unbalanced panel data with various missing values. In a general setup, we develop a multi-layer neural network to augment additional long-short latent factors to a factor model like CAPM. The approach explores the firm characteristic search space via various nonlinear transformations with an economic objective: to eliminate mispricing alphas. Our algorithm provides a joint estimation for the augmented linear factor model and the underlying neural network. To illustrate the method, we design our long-short latent factor construction in a train-validation- test framework. From an empirical perspective, we perform an out-of-sample study to analyze Fama-French factors in both the cross-section and time series. The finding is a significant forecasting improvement by adding nonlinear signals from firm characteristics.

E0319: **Volatility uncertainty and the cross-section of option returns**
Presenter: **Jie Cao**, Chinese University of Hong Kong, Hong Kong
The aim is to study how the uncertainty of volatility change predicts the cross-section of delta-hedged equity option returns. We use three estimators of time-varying daily volatility: option implied volatility, volatility of daily returns under EGARCH model, and realized volatility of intra-day returns. We find that delta-hedged option returns consistently decrease with the volatility of volatility change. The results are robust to firm characteristics, to stock and option liquidity, to volatility characteristics, to jump risks, and are not explained by standard risk factors. It suggests that option dealers charge a higher premium for options on high volatility uncertainty stocks which are more difficult to hedge.

E0358: **Attention to global warming**
Presenter: **Zhenyu Gao**, Chinese University of Hong Kong, Hong Kong
Co-authors: Wenxi Jiang, Darwin Choi
It is claimed that people update their beliefs about climate change when there are attention-grabbing weather events in their area. The effects of long-

term global warming may be overlooked in normal times, but people revise their beliefs upwards under extreme local weather conditions. Using international data, we show that attention to climate change, proxied by Google and Bloomberg search volume, goes up when the local temperature is abnormally high. In financial markets, stocks of carbon-intensive firms underperform firms with low carbon emissions in abnormally warm weather. We shed light on understanding collective beliefs and actions in response to global warming.

E0621: **Dynamic analysis of the ARP investment universe**
Presenter: **Serge Darolles**, Paris Dauphine, France
Co-authors: marie lambert
The aim is to study the persistence over time of the performance of alternative risk premiums. Using this approach, we test whether or not these premiums are associated with one of several systematic dimensions of risk. If so, their performance should not decrease when the number of financial products allowing exposure to these premiums increases (diversity of the category), or when the exchange volume also increases. We use to a new database that brings together a large number of indices marketed by banks and asset managers, and allowing a direct and liquid exposure to the numerous risk premiums identified by academics and practitioners.

EO182 Room G4701 INCOMPLETE DATA AND STATISTICS IN HEALTH STUDIES Chair: Christian Heumann

E0278: **Robust semiparametric Bayesian methods in growth curve modeling with nonnormal missing data**
Presenter: **Xin Tong**, University of Virginia, United States
Despite the wide application of growth curve models in health research, few studies have dealt with two practical issues of longitudinal data analysis – nonnormality of data and missing data. A semiparametric Bayesian approach is proposed for growth curve modeling, in which intraindividual measurement errors follow unknown distributions with Dirichlet process priors. To deal with missing data, a multiple imputation technique is applied for missing completely at random and missing at random data. A Monte Carlo simulation study is conducted to evaluate the proposed method and compare it to the traditional growth curve modeling. A real data analysis in health research is also provided to illustrate the application of the semiparametric growth curve modeling.

E0593: **Statistical methods for micro- and macro-level accelerometry data**
Presenter: **Jiawei Bai**, Johns Hopkins University, United States
Co-authors: Ciprian Crainiceanu
Wearable devices, such as accelerometers and heart rate monitors, can now provide objective and continuous measurements of human activity. Such devices have been widely deployed in large observational and clinical studies because they are expected to produce objective measurements that could improve or replace current self-reported activity measuring practices. Accelerometry data were usually obtained in a very high sampling frequency (micro-level), and could be subsequently reduced to count data (macro-level) with one measurement per minute for easier usage. Different statistical methods are needed to analyze the accelerometry signals of different level. We first introduce a movement recognition method based on movelets, to predict the type of physical activity at the sub-second level using the micro-level accelerometry data. Then, we discuss a two-stage model for the macro-level data to describe the inactive/active and activity intensity dynamics of the circadian rhythm of physical activity.

E0297: **Medical image analysis and its applications**
Presenter: **Yao Lu**, Sun Yat-Sen University, China
Medical images provide functional and structural clinical information. Its crucial and urgent to find solutions on enhancing quality of medical images and extracting disease-related information. This report tries to take breast cancer as an example to illustrate basic theories and techniques in medical image analysis and Radiomics. Recent studies and progress in high dimensional reconstruction of complicated medical data, feature extraction, feature reduction, and machine learning are also discussed.

E0485: **Nearest neighbor imputation in longitudinal studies**
Presenter: **Shahla Faisal**, Ludwig Maximilians University Munich, Germany
Co-authors: Christian Heumann
Longitudinal data often comes with missing values. These values cannot be ignored as it can result in loss of important information regarding samples. Therefore, imputation is a good strategy to overcome this problem. We present a single imputation method based on weighted nearest neighbors that uses the information from other variables to estimate the missing values. These neighbors use the information from within the sample whose response is measured at different time points and between samples. The simulation results show that the suggested imputation method provides better results with smaller imputation errors. Moreover the method performs in high dimensional data as good as in low dimensional data sets.

E0487: **Time efficient multiple imputation with penalization for high-dimensional data**
Presenter: **Faisal Maqbool Zahid**, Ludwig-Maximilians-University Munich Germany, Germany
Co-authors: Christian Heumann
The analysis of modern data based on high-throughput technology often faces the problem of missing data. Multiple imputation (MI) by sequential regression is a flexible and practical approach to handling the missing data. The precise strategy to conduct MI in the presence of high-dimensional data is still not clear in the literature. The decision about the number of predictors in the imputation model is also arguable in the literature. The likelihood estimates become unstable when the number of predictors p is large relative to the sample size n, and do not exist for $p > n$. For selection and fitting of the imputation model, we use penalization in different ways in the presence of high-dimensional data. We tune the L1 penalty to allow different number of informative predictors in the imputation model, and then use maximum likelihood estimation or L2 penalty for fitting the imputation model. We compared the performance of our proposed approaches in high-dimensional data structures through different simulation studies and two real life datasets. The proposed approach is time efficient and performs equally well in low dimension. The proposed methods show a better performance than the existing MI approaches in terms of Mean Squared Imputation Error (MSIE) and MSE($\hat{\beta}$).

EO224 Room LT-11 NONLINEARITY IN REGRESSION MODELS Chair: Feng Yao

E0166: **A four-component semiparametric stochastic frontier model with endogenous regressors and determinants of inefficiency**
Presenter: **Kai Sun**, Shanghai University, China
Co-authors: Subal Kumbhakar
A semiparametric stochastic production frontier model is proposed where the technology parameters are unknown smooth functions of environmental variables, and inputs are allowed to be endogenous. There are four components in the error term of this stochastic frontier model, where two of them are the noise components including the time-invariant and time-varying noises, and the other two of them are the inefficiency components including the time-invariant (i.e., persistent) and time-varying (i.e., transient) inefficiencies. The transient inefficiency is allowed to be a function of the environmental variables as well. We apply the proposed methodology to the Norwegian salmon production data, and analyze the estimated smooth coefficients (i.e., input elasticities), marginal effects of farm age, and persistent, transient, and overall technical efficiency scores.

E0176: **Intermediate goods price shock, vertical trade and exchange rate regime**
Presenter: **Zhouheng Wu**, Guangdong University of Foreign Studies, China
Co-authors: Kang Shi, Juanyi Xu
With the rapid growth of vertical trade in small open economies, the fluctuation of global intermediate goods price has become one of major uncertainties faced by these economies. A simple analytical two sector sticky price model is developed to show how global intermediate goods price shock affects small open economies through vertical trade. We find that welfare effects depend critically on the structure of vertical trade and exchange rate policy regime. Furthermore, we estimate an infinite period small open economy model using Bayesian Method and Canadian data. The results show that intermediate goods price shock can explain 42% of the variance of output and 68% of the variance of trade balance to output ratio, and welfare effects of intermediate goods price shock on the Canadian economy are even larger than those of technology shock. With the counterfactual experiment, we show that flexible exchange rate regime and high financial integration can mitigate the impacts of world intermediate goods price shock.

E0200: **A consistent gradient-based nonparametric test for regression structures**
Presenter: **Taining Wang**, West Virginia University, United States
Co-authors: Feng Yao
A consistent nonparametric test is considered for the relevant variables in the gradient function of the regression model, which can be used to detect the interaction among regressors and nonlinearity of a single regressor in a nonparametric regression. Our test statistics are based on the first-order gradient obtained by local quadratic estimation and we obtain its empirical distribution via bootstrap. Regarding the contribution in empirical studies, we show that it can be applied to identify regression structures, including additive, varying coefficient, and partially linear models, thereby providing statistical evidence for which semiparametric structure should be implemented in practice.

E0286: **Estimation of a smooth coefficient zero-inefficiency panel stochastic frontier model: A semiparametric approach**
Presenter: **Feng Yao**, West Virginia University, United States
Co-authors: Taining Wang, Jinjing Tian, Subal Kumbhakar
A zero-inefficiency stochastic frontier model with a simple semiparametric approach using panel data is proposed. We model the frontier with a smooth coefficient function and specify a nonzero conditional probability for firms being fully efficient to be a known function of environment variables. Following previous work, we propose a three step semiparametric estimator which is computationally convenient. The simulation results reveal encouraging finite sample properties. We illustrate the applicability of our model using country level data from the Penn World Table.

E0687: **Monetary shock measurement and stock markets**
Presenter: **Arabinda Basistha**, West Virginia University, United States
Co-authors: Richard Startz
Narrative approach based measurement of monetary shocks suggest infrequent shocks are crucial for understanding the impact of monetary policy shocks on the economy. However, the narrative approach is also dependent on costly data collection process, researcher judgement and prone to delays due to official document release. We present a stock market based non-linear empirical model to estimate monetary shocks while preserving the key feature of infrequent shocks. Our estimated shocks are large and comparable to previous ones. The estimated impulse responses suggest that a one percent contractionary shock leads to two percent long term decline in industrial production with a peak effect of 3.5 percent decline and more than one percent long term decline in CPI.

EO149 Room LT-12 STATISTICAL MACHINE LEARNING Chair: Yiming Ying

E0175: **A comparison of matching and machine learning-based covariate adjustment**
Presenter: **Luke Keele**, Georgetown University, United States
Co-authors: Dylan Small
Matching algorithms have become one frequently used method for statistical adjustment under a selection on observables identification strategy. Matching methods typically focus on modeling the treatment assignment process rather than the outcome. Many of the recent advances in matching allow for various forms of covariate prioritization. This allows analysts to emphasize the adjustment of some covariates over others, typically based on subject matter expertise. While flexible machine learning methods have a long history of being used for statistical prediction, they have generally seen little use in causal modeling. However, recent work has developed flexible machine learning methods based on outcome models for the estimation of causal effects. These methods are designed to use little analyst input. All covariate prioritization is done by the learner. In this study, we replicate five published studies that used customized matching methods for covariate prioritization. In each of these studies, subsets of covariates were given priority in the match based on substantive expertise. We replicate these studies using BART, a machine learning method that has been used for causal modeling. We record differences in both point estimates, confidence interval length, and sample trimming.

E0197: **Convergence of gradient descent method for minimum error entropy principle**
Presenter: **Ting Hu**, Wuhan University, The Hong Kong Polytechnic University, China
Information theoretical learning refers to a framework of learning methods that use concepts of entropies and divergences from information theory to substitute the conventional statistical descriptors of variances and covariance. It becomes an important research topic in signal processing and machine learning as many algorithms have been developed within this framework and many applications domains have been discovered. We study a kernel version of minimum error entropy methods that can be used to find non-linear structures in the data. We show that the kernel minimum error entropy can be implemented by kernel based gradient descent algorithms with or without regularization.

E0205: **Online learning for supervised dimension reduction**
Presenter: **Qiang Wu**, Middle Tennessee State University, United States
Co-authors: Ning Zhang
Supervised dimension reduction is an effective tool for high dimension data analysis. It enables easy visualization of the data and improves predictive power of subsequent analyses by other statistical machine learning algorithms. As high dimensional and big data become ubiquitous in modern sciences, it is necessary to develop fast and dynamic supervised dimension reduction methods. We will present two new methods that implement dimension reduction in an online learning manner. These methods are much faster than batch learning methods while achieve comparable performance.

E0260: **Spectral algorithms for functional linear regression**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong
Functional data analysis is concerned with inherently infinite dimensional data such as curves or images. It attracts more and more attentions because of its successful applications in many areas such as neuroscience and econometrics. We consider a class of regularization methods called spectral algorithms for functional linear regression within the framework of reproducing kernel Hilbert space. The proposed estimators can achieve

the minimax optimal rates of convergence. Despite of the infinite dimensional nature of functional data, we show that the algorithms are easily implementable.

E0340: **Reproducing kernels for pairwise learning**
Presenter: **Xin Guo**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Ting Hu, Qiang Wu
Pairwise learning is a large family of learning algorithms for the problems where supervised labels are not available, but one has only the access to the differences between labels of each pair of sample points. For example, ranking, AUC maximization, metric learning, gradient learning, and so on. We studied a transform of reproducing kernels so that the obtained kernels fit the purpose of pairwise learning way better. The relation between the integral operators and the hypothesis spaces of the original and the transformed kernels are obtained.

EO328 Room LT-13 FINANCIAL ECONOMETRICS Chair: Sandra Paterlini

E0699: **Creating (parsimonious) banking networks**
Presenter: **Sandra Paterlini**, European Business School Germany, Germany
Co-authors: Dietmar Maringer, Ben Craig
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.

E0786: **Consumption-based risk of bonds and stocks**
Presenter: **Svetlana Bryzgalova**, Stanford Graduate School of Business, United States
Co-authors: Christian Julliard
Aggregate consumption growth reacts slowly, but significantly, to bond and stock return innovations. As a consequence, slow consumption adjustment (SCA) risk, measured by the reaction of consumption growth cumulated over many quarters following a return, can explain most of the cross-sectional variation of expected bond and stock returns. Moreover, SCA shocks explain about a quarter of the time series variation of consumption growth, a large part of the time series variation of stock returns, and a significant (but small) fraction of the time series variation of bond returns, and have substantial predictive power for future consumption growth.

E0750: **Some moment problems arising in financial econometrics**
Presenter: **Christian Kleiber**, Universitaet Basel, Switzerland
The moment problem asks whether a distribution can be uniquely characterised by the sequence of its moments. Distributions that are not characterised by the sequence of their moments have long been known, e.g., the lognormal and certain generalised gamma distributions. We consider models from financial econometrics in which moment-indeterminate distributions may arise. Specifically, the generalised error distribution (GED) appearing in the EGARCH model is moment-indeterminate for some values of the parameters. Also, we show that one of the standard volatility models in financial econometrics, namely the stochastic volatility (SV) model, leads to return distributions that are moment-indeterminate. Perhaps somewhat unexpectedly, moment indeterminacy already arises in the classical discrete time SV model with lognormal latent volatility and independent multiplicative Gaussian noise.

E0469: **Trending heterogeneous and time-varying coefficient panel data models with endogeneity and fixed effects**
Presenter: **Li Chen**, Xiamen University, China
The aim is to study a trending heterogeneous and time-varying coefficient panel data model with endogeneity and fixed effects. The regression coefficients change over both time and sections. We propose a projection method to solve the problem of endogeneity. Meanwhile, we employ an easy-to-implement series expansion method to estimate the coefficients. We show that the estimators are consistent as the cross-sectional size N and time series length T tend to infinity. We also investigate the relationship between GDP and healthcare expenditure in OECD countries as an empirical example.

E0183: **Nonparametric inference on the self-excitation of jumps in jump diffusion models**
Presenter: **Simon Kwok**, University of Sydney, Australia
Understanding the jump dynamics of market prices is important for derivative pricing and risk management. Despite their analytical tractability, parametric jump diffusion models entail restrictive and unrealistic structure on the jump dynamics. We propose a set of nonparametric estimator for jump autocorrelation associated with different powers of the log-return process. The nonparametric estimator is consistent for the jump autocorrelation measure and asymptotically normal under mild moment and stationarity conditions. This enables pointwise inference through the construction of jump auto-correlogram with confidence bounds. Furthermore, we study an omnibus test for no self-excitation of jumps at all positive lag orders. The method is naturally extendable to the inference of cross-correlation of jumps in a bivariate setting. In an empirical study of jump contagion in stock markets, we found richer jump dynamic structure that is different from what was implied from conventional jump diffusion models in the literature.

EO034 Room LT-14 EFFICIENT LEARNING FOR LARGE-SCALE DATA Chair: Wei Zheng

E0653: **High-dimensional Gaussian graphical model for network-linked data**
Presenter: **Ji Zhu**, University of Michigan, United States
Graphical models are commonly used in representing conditional independence between random variables, and learning the conditional independence structure from data has attracted much attention in recent years. However, almost all commonly used graph learning methods rely on the assumption that the observations share the same mean vector. We extend the Gaussian graphical model to the setting where the observations are connected by a network and propose a model that allows the mean vectors for different observations to be different. We have developed an efficient estimation method for the model and demonstrated the effectiveness of the proposed method using simulation studies. Further, we prove that under the assumption of "network cohesion", the proposed method can estimate both the inverse covariance matrix and the corresponding graph structure accurately. We have also applied the proposed method to a dataset consisting of statisticians' coauthorship network to learn the statistical term dependency based on the authors' publications and obtained meaningful results.

E0303: **Asympirical method: A new paradigm for the statistical analysis of large samples**
Presenter: **Ping Ma**, University of Georgia, United States
Traditional statistical theory and methods are developed for small and mild size samples. In particular, statistical model fitting and inference are conducted in the small and mild size samples to get empirical results. Asymptotic theory is established to extrapolate the performance of the

empirical results to large samples. However, this traditional coherent statistical analysis paradigm falls apart in large samples. The key challenge is that many traditional statistical methods are computational too expensive to get meaningful empirical results. A new statistical paradigm is in urgent need for the statistical analysis in large samples. We will present an asympirical (asymptotic + empirical) method, which is designed by the principle that theory informs practice. We will present it in the context of smoothing spline ANOVA models. Simulation and real data analysis will be used to demonstrate the performance of the new paradigm.

E0209: **Enabling phenotypic big data with PheNorm**
Presenter: **Sheng Yu**, Tsinghua University, China
Co-authors: Tianxi Cai
EHR-based phenotyping infers whether a patient has a disease based on the information in their electronic health records (EHR). A human annotated training set with gold-standard disease status labels is usually required to build an algorithm for phenotyping based on a set of predictive features. The time intensiveness of annotation as well as feature curation severely limits the ability to achieve high-throughput phenotyping. While previous studies have successfully automated feature curation, annotation remains a major bottleneck. We present PheNorm, a phenotyping algorithm that does not require expert-labeled samples for training. PheNorm transforms predictive features, such as the number of ICD-9 codes or mentions of the target phenotype, to resemble a normal mixture distribution. The transformed features are then denoised and combined into a score for accurate disease classification. We validated the accuracy of PheNorm with four phenotypes: coronary artery disease, rheumatoid arthritis, Crohns disease, and ulcerative colitis. The AUC of the PheNorm score reached 0.90, 0.94, 0.95, and 0.94 for the four phenotypes, respectively, which were comparable to the accuracy of supervised algorithms trained with sample sizes of 100-300, with no statistically significant difference.

E0453: **Global testing under sparse alternative for single index model**
Presenter: **Qian Lin**, Tsinghua University, China
Testing for the significance of a signal in a linear model goes back at least to the work of Fisher. We study this problem for the single index model $y = f(\beta^\tau x, \varepsilon)$ with Gaussian design where f is unknown and β is a p dimensional unit vector with at most s nonzero entries. We adopt the notion of generalized signal to noise ratio (gSNR). We are interested in the hypothesis testing problem of whether $\beta = 0$ or not. Let n be the size of observed data. We show that if $s^2 \wedge p \prec n$, one can detect the gSNR if and only if $gSNR \succ \frac{p^{1/2}}{n} \wedge \frac{s\log(p)}{n}$. Furthermore, if the noise is additive(i.e., $y = f(\beta^\tau x) + \varepsilon$), one can detect gSNR if and only if $gSNR \succ \frac{p^{1/2}}{n} \wedge \frac{s\log(p)}{n} \wedge \frac{1}{\sqrt{n}}$. In other words, the detection boundary gSNR for the single index model with additive noise matches that of SNR for linear regression. These results pave the road of a through treatment of single/multiple index models in high dimensions. For example, one may try to extend the well developed theories of linear models to the single/multiple index models with Gaussian design.

E0797: **An asymptotically efficient test for functional coefficient models**
Presenter: **Xingtong Zhang**, Cornell University, United States
Functional coefficient models have abilities to capture non-linearity and heteroscedasticity by allowing coefficients to be governed by some variables. To test the model specifications, there are two general approaches: the generalized likelihood ratio (GLR) test proposed and loss function approach. Despite enjoying appealing features such as Wilks phenomena, they both rely on nonparametric convergence rate and thus suffer from curse of dimensionality. We propose a root-T consistent test using Fourier transforms. The new test is asymptotically more efficient than both GLR test and loss function approach. Because of its parametric convergence rate, our test is free from dimension of nonparametric smoothing.

EO192 Room LT-16 RECENT ADVANCES IN BAYESIAN NONPARAMETRIC THEORY Chair: Botond Szabo

E0186: **Nonparametric Bayesian analysis for support boundary recovery**
Presenter: **Johannes Schmidt-Hieber**, Leiden University, Netherlands
Co-authors: Markus Reiss
Frequentist properties of the posterior distribution for a boundary detection problem are investigated. More specifically, given a sample of a Poisson point process with positive intensity above a boundary function f and zero intensity below the boundary function, we study recovery of f from a nonparametric Bayes perspective. Because of the irregularity of this model, the analysis is non-standard. We derive contraction rates for several classes of priors, including Gaussian priors, priors based on (truncated) random series, compound Poisson processes, and subordinators. We also investigate the limiting shape of the posterior distribution and derive a nonparametric version of the Bernstein-von Mises theorem for a specific class of priors on a function space with increasing parameter dimension. We show that the marginal posterior of the integral over f does some automatic bias correction and contracts with a faster rate than the MLE. In this case, credible sets are also asymptotic confidence intervals. It is also shown that the frequentist coverage of credible sets is lost under model misspecification.

E0206: **Convergence rates of variational posterior distributions**
Presenter: **Chao Gao**, University of Chicago, United States
Convergence rates of variational posterior distributions for nonparametric and high-dimensional inference are studies. We formulate general conditions on prior, likelihood, and variational class that characterize the convergence rates. Under similar "prior mass and testing" conditions considered in the literature, the rate is found to be the sum of two terms. The first term stands for the convergence rate of the true posterior distribution, and the second term is contributed by the variational approximation error. For a class of priors that admit the structure of a mixture of product measures, we propose a novel prior mass condition, under which the variational approximation error of the generalized mean-field class is dominated by convergence rate of the true posterior. We demonstrate the applicability of our general results for various models, prior distributions and variational classes by deriving convergence rates of the corresponding variational posteriors.

E0356: **Posterior concentration for Bayesian regression trees and their ensembles**
Presenter: **Stephanie van der Pas**, Leiden University, Netherlands
Co-authors: Veronika Rockova
Since their inception in the 1980s, regression trees have been one of the more widely used nonparametric 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 overfitting. Strategies against overfitting have traditionally relied on pruning greedily grown trees. The Bayesian framework offers an alternative remedy against overfitting through priors. Roughly speaking, a good prior charges smaller trees where overfitting does not occur. We take a step towards understanding why/when do Bayesian trees and their ensembles not overfit. 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.

E0491: **Coverage aspects of Gaussian processes with an application to particle Physics**
Presenter: **Debdeep Pati**, Texas A&M University, United States
Co-authors: Anirban Bhattacharya, Yun Yang
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. 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. 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. We demonstrate the validity of our theoretical results through numerical examples and an application to the famous proton radius puzzle.

E0573: **Nonparametric Bayesian contraction rates for compound Poisson processes observed discretely at arbitrary frequencies**
Presenter: **Alberto J Coca**, University of Cambridge, United Kingdom
Compound Poisson processes (CPPs) are the textbook example of pure jump stochastic processes, and they approximate arbitrarily well much richer classes of processes such as Lévy processes. Two parameters characterise them: the drift, and the Lévy jump distribution, N, driving the frequency at which jumps (randomly) occur and their (random) sizes. In most applications, the CPP is not perfectly observed: only discrete observations over a finite-time interval are available. Thus, the process may jump several times between two observations and estimating N is a non-linear statistical inverse problem. In the recent years, understanding the asymptotic behaviour of the Bayesian method in inverse problems and, in particular, in this problem has received considerable attention. We will present some recent results on posterior contraction rates for the density ν of N: we show two-sided stability estimates between ν and its image through the forward operator that allow to use existing classical theory; furthermore, these are robust to the observation interval, i.e. optimal adaptive inference can be made without specification of whether the regime is of high- or low-frequency; and, lastly, we propose an efficient ∞-MCMC procedure to draw from the posterior using mixture and Gaussian priors that can handle the multidimensional setting.

EO032 Room LT-18 RECENT DEVELOPMENT FOR MODERN CHANGE-POINT ANALYSIS Chair: Hao Chen

E0441: **Segmentation and estimation of change-point models**
Presenter: **Xiao Fang**, The Chinese University of Hong Kong, Hong Kong
To segment a sequence of independent random variables at an unknown number of change-points, new procedures are introduced that are based on thresholding the likelihood ratio statistic. We also study confidence regions based on the likelihood ratio statistic for the change points and joint confidence regions for the change-points and the parameter values. Applications to segment an array CGH analysis of the BT474 cell line are discussed.

E0576: **A general theory for detecting changes-in-mean and changes-in-slope**
Presenter: **Chao Zheng**, Lancaster University, United Kingdom
Co-authors: Idris Eckley, Paul Fearnhead
The aim is to study the finite sample behaviour of an approach to detecting changepoints that is based on maximising a penalised likelihood. These give general results as to when such a procedure can consistently estimate the number of changes and accurately estimate their position. The results we obtained are applied to the problem of detecting changes-in-mean and changes-in-slope. In the latter case we obtain tighter results on the value of penalty that can be used as compared to existing theory. Moreover, the techniques can be easily adapted to other scenarios as long as some basic properties for detecting a single changepoint are satisfied. We demonstrate the usefulness of our approach through numerical experiments on both synthetic data and real data examples.

E0652: **A super scalable algorithm for short segment detection**
Presenter: **Yue Niu**, University of Arizona, United States
Co-authors: Ning Hao, Heping Zhang, Feifei Xiao
In many applications such as Copy Number Variation (CNV) detection, the goal is to identify short segments which have means different from the background. The target segments are usually short and hidden in a long sequence, which make the problem very challenging to solve. We introduce a super scalable short segment detection algorithm, which is fast in computation. It is a nonparametric method that does not rely on Gaussian assumption. Moreover, it can assign a significance level for each of detected segment.

E0795: **Consistent selection of the number of change-points via sample-splitting**
Presenter: **Guanghui Wang**, Nankai University, China
In multiple change-point analysis, one of the major challenges is to estimate the number of change-points. Most existing approaches attempt to minimize a Schwarz information criterion which balances a term quantifying model fit with a penalization term accounting for model complexity that increases with the number of change-points and limits overfitting. However, different penalization terms are required to adapt to different contexts of multiple change-point problems and the optimal penalization magnitude usually varies from the model and error distribution. We propose a data-driven selection criterion that is applicable to most kinds of popular change-point detection methods, including binary segmentation and optimal partitioning algorithms. The key idea is to select the number of change-points that minimizes the squared prediction error, which measures the fit of a specified model for a new sample. We develop a cross-validation estimation scheme based on an order-preserved sample-splitting strategy, and establish its asymptotic selection consistency under some mild conditions. The proposed selection criterion's effectiveness is demonstrated on a variety of numerical experiments and real-data examples.

E0520: **Change point detection in high-dimensional time series with both spatial and temporal dependence**
Presenter: **Jun Li**, Kent State University, United States
High-dimensional time series are characterized by a large number of measurements and complex dependence, and often involve abrupt changes at unknown time points. We will present a new procedure to detect the change points from high-dimensional time series data. An asymptotic testing procedure is established for the hypothesis of existing any change point. When the null hypothesis is rejected, a binary segmentation method is conducted to estimate multiple change points. We will demonstrate the impact of sample size, dimensionality and the location of the change point on the proposed method. Compared with other methods, the proposed procedure allows both sample size and dimensionality to diverge without constraint on the growth rate of dimensionality. Moreover, it does not assume Gaussianity, and incorporates both spatial and temporal dependence without imposing restrictive structural assumptions.

EO044 Room P4701 STATISTICAL METHODS FOR SYSTEMS MONITORING Chair: Inez Zwetsloot

E0343: **A transfer learning approach for modeling and monitoring in landslide sensor systems**
Presenter: **Ke Zhang**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Zhenli Song, Fugee Tsung
Landslides are common geographical activities that result in large quantities of rock and debris flowing down hill-slopes, leading to thousands of casualties and billions of dollars in infrastructure damage every year around the world. To detect such abnormal geographical behavior, on-site sensor systems are widely applied for data collection and many existing SPC methods can be adopted for modeling and monitoring. However, the conventional methods may fail to perform well for newly set-up sensors with small data collected. To make use of the new sensors effectively right after any scale-up of the system, we proposed a transfer learning based approach to jointly model the sensor data streams thus getting better understanding on new sensors by the information transferred from old sensors. In the approach, the parameters within auto-regressive models for individual sensors are connected using a Gaussian prior and certain regularization terms. An iterative updating scheme has been developed for parameter estimation in the integrated model. After modeling, a residual-based monitoring scheme is proposed accordingly. Various Monte Carlo simulations have been conducted to illustrate the performance of our transfer learning method over conventional ones. Real data example also shows that the proposed method can be effectively applied in real landslide monitoring applications.

E0399: **Variable sampling interval np chart with estimated parameter**
Presenter: **Shu Wu**, Wuhan University of Technology, China
In recent years, the variable sampling interval(VSI) X-bar chart has been investigated by many researchers, because the VSI control charts have important application and useful in service industries. The VSI control charts sample at a higher rate when there is evidence of a change in the process, and are thus able to detect process changes faster than traditional control charts. But the properties of the attribute control charts like VSI np chart with estimated parameter is not investigated yet. The performances of the VSI np chart are evaluated and compared in terms of its average time to signal(ATS), in the case where the process parameter is known and is estimated, the results demonstrate that the performances of VSI np chart are quite different when the number of samples used during Phase I is small. The number of Phase I samples are computed to provide a new optimal constants to have approximately the same in-control average time to signal. An optimization technique to find the suitable chart parameter of VSI np chart is also developed.

E0405: **A distribution-free multivariate change-point model for statistical process control**
Presenter: **Maoyuan Zhou**, Civil Aviation University of China, China
A new distribution-free multivariate procedure is developed for statistical process control based on minimal spanning tree (MST), which integrates a multivariate two-sample goodness-of-fit (GOF) test based on MST and change-point model. Simulation results show that our proposed procedure is quite robust to nonnormally distributed data, and moreover, it is efficient in detecting process shifts, especially moderate to large shifts, which is one of the main drawbacks of most distribution-free procedures in the literature. The proposed procedure is particularly useful in start-up situations. Comparison results and a real data example show that our proposed procedure has great potential for application.

E0642: **Generalized design of control chart for weighted-count data under measurement error**
Presenter: **Wichai Chattinnawat**, Chianf Mai University, Thailand
Consider an online automated high speed quality monitoring system where the process quality is determined by the number and types of nonconformity. In general, Demerit control chart can be adopted to monitor the system where the score calculated from the quantity, type, and severity of the nonconformities is used as control chart statistic. The process is considered acceptable as long as the calculated demerit score remains within present limits. However, there is no systematic investigation of optimum strategy presented in the literature on how to design the demerit weights of general types by taking into account the distribution of the demerit score which is not normally distributed as imposed by the traditional demerit control chart assumption. Moreover, the online high speed quality measurement such as based pattern recognition technology can be associated with some degree of measurement errors. Putting altogether the traditional demerit control chart needs to be studied and generalized. A generalized design methodology is presented for weighted-count type control chart with arbitrary weights structure to monitor several types of nonconformity when the nonconformity types are independent but possibly masked with measurement error. The aim is to propose a design methodology to derive the optimum weights and a new design methodology of self adapting weights.

E0460: **Monitoring data quality in a personalized health tracking system**
Presenter: **Inez Zwetsloot**, City University of Hong Kong, Hong Kong
Rapid advances in information and sensor technology have led to the development of tools and methods for individual health monitoring. These techniques support elderly health management by tracking vital signs and detecting physiological changes. We develop a data quality monitoring system to ensure that the collected data is accurate. We consider data from an all-in-one station-based health monitoring device that collects daily vital signs of elderly in an elderly home in Hong Kong. Due to the nature of the data both the sample sizes as well as the number of measured variables changes over time. We develop a new control chart to monitor the data quality. This new method has the ability to monitor data with varying sample sizes and varying number of parameters effectively. We illustrate the new method using the data on vital signs of the participants.

EO018 Room P4703 ESTIMATION, MODELING CHECKING, AND DIMENSION REDUCTION Chair: Sung Nok Chiu

E0401: **A weighted learning approach for sufficient dimension reduction in binary classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South
In binary classification, sufficient dimension reduction (SDR) often suffers from the lower resolution of binary responses. For example, the sliced inverse regression can estimate at most one basis of the central subspace. A new class of SDR algorithm in binary classification is proposed based on weighted learning. Toward this, we establish that the gradient of the decision function is unbiased for SDR if the loss function of the classifier is Fisher consistent. This naturally leads us to develop a corresponding working matrix whose first few eigenvectors estimate the basis set of the central space for the binary response. The performance of the proposed method is evaluated by both simulated and real data examples.

E0678: **A constrained maximum likelihood estimation for skew normal mixtures**
Presenter: **Libin Jin**, Shanghai Lixin University of Accounting and Finance, China
Co-authors: Sung Nok Chiu, Lixing Zhu
For a finite mixture of skew normal distributions, the maximum likelihood estimator is not well-defined because of the unboundedness of the likelihood function when scale parameters go to zero and the divergency of the skewness parameter estimates. To overcome these two problems simultaneously, we propose constrained maximum likelihood estimators under constraints on both the scale parameters and the skewness parameters. The proposed estimators are consistent and asymptotically efficient under relaxed constraints on the scale and skewness parameters. Numerical simulations show that in finite sample cases the proposed estimators outperform the ordinary maximum likelihood estimators. A real data set of Iris flowers is used to illustrate the success of the proposed approach.

E0679: **A method selection guidance in dr-package**
Presenter: **Jae Keun Yoo**, Ewha Womans University, Korea, South

Sufficient dimension reduction (SDR) turns out to be a useful dimension reduction tool in high-dimensional regression analysis. The dr-package has been developed to implement the four most popular SDR methods. However, the package does not provide any clear guideline on which method should be used given a data. Since the four methods may provide dramatically different dimension reduction results, the selection in dr package should be problematic to statistical practitioners. A basis-adaptive selection algorithm is developed to relieve the issue. A basic idea is to select a SDR method to provide a highest correlation between the basis estimates obtained by the four classical SDR methods. Real data example and numerical studies confirm practical usefulness of the algorithm.

E0681: **Adaptive model checking for functional single-index models**
Presenter: **Qing Jiang**, Beijing Normal University, China
Co-authors: Feifei Chen, Zhenghui Feng

Functional model checking is seldom studied in the literature because of the infinite dimension trouble. A model-adaptive test statistic is proposed to do model checking for functional single-index models. Dimension reduction methods are included to handle the curse of dimensionality. The test consists of two parts: the first term is a naive one, and the second term is adaptive to the model as if the model were univariate, and with higher power. It is consistent against any global alternative hypothesis and can detect local alternative at a fast rate. Simulation studies show the performance compared with other methods.

E0683: **Simple structure estimation via prenet penalization in factor analysis model**
Presenter: **Kei Hirose**, Kyushu University, Japan
Co-authors: Yoshikazu Terada

A prenet (product elastic net) is proposed, which is a new penalization method for factor analysis models. The penalty is based on the product of a pair of elements in each row of the loading matrix. The prenet not only shrinks some of the factor loadings toward exactly zero, but also enhances the simplicity of the loading matrix, which plays an important role in the interpretation of the common factors. In particular, with a large amount of prenet penalization, the estimated loading matrix possesses a perfect simple structure, which is known as a desirable structure in terms of the simplicity of the loading matrix. Furthermore, the perfect simple structure estimation via the prenet turns out to be a generalization of the k-means clustering of variables. On the other hand, a mild amount of the penalization approximates a loading matrix estimated by the quartimin rotation, one of the most commonly used oblique rotation techniques. Thus, the proposed penalty bridges a gap between the perfect simple structure and the quartimin rotation. Some real data analyses are given to illustrate the usefulness of our penalization.

EO022 Room P4704 RECENT ADVANCES IN SURVIVAL ANALYSIS Chair: Xingqiu Zhao

E0597: **Exact inference of point and interval estimation for Laplace distribution based on various censoring**
Presenter: **Xiaojun Zhu**, Xi'an Jiaotong-Liverpool University, China

Some recent results of the exact inference of the Laplace distribution based on various censored samples will be presented. Under these censoring forms, it can be shown that the maximum likelihood estimators, best linear unbiased estimators as well as the predictive maximum likelihood estimators are all in the form of linear combinations of order statistics from Laplace distribution. Then, it is possible to construct the exact inference for these estimators and predictors. Using the same idea, the exact confidence intervals for the estimation of survival and cumulative hazard functions can be obtained.

E0635: **Inverse probability weighting methods for the analysis of panel count data with informative observation times**
Presenter: **Ni Li**, School of Mathematics and Statistics, Hainan Normal University, China

Recurrent event data usually occur in long-term studies which concern recurrence rates of certain events. In some circumstances of these studies, subjects can only be observed at discrete time points rather than continuously and thus only the numbers of the events that occur between the observation times, not their occurrence times, are observed. This type of data can also be referred to as interval-censored recurrent event data, or panel count data. In panel count data, the observation times or process may differ from subject to subject and more importantly, may contain relevant information about the underlying recurrent event process, therefore can be viewed as dependent observation process. Methods have been proposed for regression analysis of panel count data, but most of the existing research focuses on situations where observation times are independent of longitudinal response variables given covariates. However, the independence assumption may not hold. We propose semiparametric analysis of panel count data with adjusting for confounding effects caused by dependent observation process. In our approach, the observation filtration will be adjusted by parametric estimates of propensity scores using the idea of inverse probability weighting, to avoid confounding bias produced. The results of this research could serve as new methodologies for analyzing panel count data with informative observation times.

E0664: **A new approach to variable selection in linear mixed effect models via broken adaptive ridge regression**
Presenter: **Hong Yin**, Renmin University of China, China
Co-authors: Gang Li

Linear mixed effect models (LMEM) play an important role in the analysis of longitudinal data, panel data and cross-sectional data. They are widely used by various fields of social sciences, medical and biological sciences. However, the complex nature of these models has made variable selection and parameter estimation a challenging problem. The selection and estimation of the fixed and random effects in LMEM are considered. We propose a new variable selection method for LMEM, named broken adaptive ridge (BAR) regression which incorporates the merits of L0 penalized regression with those of ridge regression. Definitely, it is an iterative version of ridge regression in which each coefficient will be given a updated weighted score related to the last coefficients values. Due to inheriting some properties of L0-penalized regression in a sense that it can choose the non-zero components and shrink the zero components quickly, accurately and unhesitatingly. At the same time, it reserves the version of ridge regression, so there is no much burden in the optimization of objective function. We also show that the proposed method is consistent variable selection procedure and possesses some oracle properties. The results of simulation data sets and a real data have shown the efficacy of our method compared with the smoothly clipped absolute deviation penalty (SCAD).

E0776: **Functional Cox model**
Presenter: **Kin Yat Liu**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Xingqiu Zhao, Meiling Hao

Right-censored data in the presence of both functional and scalar covariates often occur in many biomedical studies. We apply a roughness regularization approach in making nonparametric inference for functional Cox models. In a reproducing kernel Hilbert space framework, we construct asymptotically confidence intervals for functional and scalar covariates and two statistical procedures for hypothesis testing with integro-differential equation techniques and functional Bahadur representation. The finite sample performance of the proposed inference procedures are evaluated through simulation studies. The proposed method is illustrated with a study of association between daily measures of the Intensive Care Unit (ICU) Sequential Organ Failure Assessment (SOFA) score and mortality.

E0802: **A novel method to generate correlated multivariate survival data with a given correlation matrix**

Presenter: **Changchun Xie**, University of Cincinnati, United States

Most financial data, clinical trials and public health studies nowadays generate multiple-endpoints, such as multivariate survival data. For simulations, the typical methods for generating correlated multivariate survival data are based on copula functions or frailty models. However, the correlation matrix in the correlated survival data generated by these methods is not very straightforward. It is not easy to use these methods to generate correlated survival data with a given correlation matrix. The aim is to propose a novel algorithm to generate correlated survival data for a given correlation matrix.

Tuesday 19.06.2018 14:00 - 15:40 Parallel Session C – EcoSta2018

EO143 Room B4302 STATISTICAL COMPUTATION FOR HIGH-DIMENSIONAL DATA AND ITS APPLICATION Chair: Kei Hirose

E0415: **A one-stage estimation of principal component regression for generalized linear models**
Presenter: **Shuichi Kawano**, The University of Electro-Communications, Japan
Co-authors: Hironori Fujisawa, Toyoyuki Takada, Toshihiko Shiroishi
Principal component regression (PCR) is a two-stage procedure: principal component analysis (PCA) is performed, and then a regression model with the selected principal components is constructed. Since PCA is based only on the explanatory variables, the principal components do not include the information on the response variable. To address the problem, we propose a one-stage estimation of PCR in the framework of generalized linear models. The loss function is based on a combination of the PCA loss and the negative log-likelihood function for an exponential family. An estimate of the parameters is obtained as the minimizer of the loss function with an L1-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. We examine the effectiveness of SPCR-glm through numerical studies.

E0618: **A component-based approach for the clustering of multivariate categorical data**
Presenter: **Michio Yamamoto**, Okayama University, Japan
A novel model-based clustering procedure for multivariate categorical data is proposed. The proposed model assumes that each response probability has a low-dimensional representation of the cluster structure in the formulation of latent class analysis. This representation, which is constructed by weights for categorical variables and component scores for cluster representatives, allows us to interpret the latent cluster structure in the categorical data. In addition, we define low-dimensional scores for individuals as convex combinations of scores for cluster representatives. It is shown that the relation between the individual scores and response probabilities can be interpreted through a divergence measure. An expectation-maximization (EM) algorithm with gradient projection and coordinate descent is developed, and it is shown that there is trade-off relation between the convergence rate of the algorithm and the cluster recovery. The usefulness of the proposed model is shown by the analysis of molecular biology data.

E0698: **High dimensional covariance matrix estimation by penalizing the matrix-logarithm transformed likelihood**
Presenter: **Xiaohang Wang**, BNU-HKBU United International College, China
It is well known that when the dimension of the data becomes very large, the sample covariance matrix S will not be a good estimator of the population covariance matrix Σ. Using such estimator, one typical consequence is that the estimated eigenvalues from S will be distorted. Many existing methods tried to solve the problem, and examples of which include regularizing Σ by thresholding or banding. We estimate Σ by maximizing the likelihood using a new penalization on the matrix logarithm of Σ (denoted by A) of the form: $\|A - mI\|_F^2 = \sum_i (\log(d_i) - m)^2$, where d_i is the ith eigenvalue of Σ. This penalty aims at shrinking the estimated eigenvalues of A toward the mean eigenvalue m. The merits of our method are that it guarantees Σ to be non-negative definite and is computational efficient. The simulation study and applications on portfolio optimization and classification of genomic data show that the proposed method outperforms existing methods.

E0543: **Data assimilation for massive autonomous systems based on a second-order adjoint method**
Presenter: **Hiromichi Nagao**, The University of Tokyo, Japan
Co-authors: Shin-ichi Ito
Data assimilation (DA) is a fundamental computational technique that integrates numerical simulation models and observation data based on Bayesian statistics. We propose an adjoint-based DA method for massive autonomous models that produces optimum estimates and their uncertainties within reasonable computation time and resource constraints. The uncertainties are given as several diagonal elements of an inverse Hessian matrix, which is the covariance matrix of a normal distribution that approximates the target posterior probability density function in the neighborhood of the optimum. Conventional algorithms for deriving the inverse Hessian matrix require $O(CN^2 + N^3)$ computations and $O(N^2)$ memory, where N is the number of degrees of freedom of a given autonomous system and C is the number of computations needed to simulate time series of suitable length. The proposed method using a second-order adjoint method allows us to directly evaluate the diagonal elements of the inverse Hessian matrix without computing all of its elements. This drastically reduces the number of computations to $O(C)$ and the amount of memory to $O(N)$ for each diagonal element. The proposed method is validated through numerical tests using a massive two-dimensional Kobayashi phase-field model. We confirm that the proposed method correctly reproduces the parameter and initial state assumed in advance, and successfully evaluates the uncertainty of the parameter.

EO109 Room G4302 RECENT ADVANCES IN TIME SERIES AND SPATIAL ANALYSIS Chair: Lily Wang

E0208: **Achieving parsimony in Bayesian VARs with the horseshoe prior**
Presenter: **Cindy Yu**, Iowa State University, United States
In the context of a vector autoregression (VAR) model, or any multivariate regression model, a large information set may be available from which to build a prediction equation. It is well known that forecasts based off of (un-penalized) least squares estimates can overfit the data and lead to poor predictions. Since the 1980's when the Minnesota prior was proposed, there have been many methods developed aiming at improving prediction performance. We propose the horseshoe prior in the context of a Bayesian VAR. The horseshoe prior is a unique shrinkage prior scheme in that shrinks irrelevant signals rigorously to 0 while allowing large signals to remain large and practically unshrunk. In an empirical study, we show that the horseshoe prior competes favorably with shrinkage schemes commonly used in Bayesian VAR models as well as with a prior that imposes true sparsity in the coefficient vector. Additionally, we propose the use of particle Gibbs with backwards simulation for the estimation of the time-varying volatility parameters.

E0225: **Composite likelihood inference for replications of spatial ordinal data**
Presenter: **Pingping Wang**, Nanjing University of Finance and Economics, China
Co-authors: Jun Zhu
Spatial ordinal data observed on multiple subjects are common in practice yet statistical methodology for such ordinal data analysis is limited. The existing methodology often assumes a single realization of spatial ordinal data without replications and thus, it is not directly applicable for the kind of spatial ordinal data observed on multiple subjects. We develop a spatial ordinal probit model that enables the assessment of covariates via regression and accounts for spatial correlation via a geostatistical model. We then develop maximum composite likelihood method for parameter estimation and establish the asymptotic properties of the parameter estimates, which differs from the existing literature in that the number of subjects tends to infinity but not the number of spatial locations per subject. The asymptotic properties permit an approximate estimation of the variance of the parameter estimates and facilitate the inference for the model parameters in a computationally efficient manner. A simulation study suggests sound finite sample properties of the proposed methods and a real data example in dental health is presented for illustration.

E0277: **Variance change point detection for data on a surface**
Presenter: **Pang Du**, Virginia Tech, United States
Co-authors: Zhenguo Gao
Motivated from an organ procurement application, we consider the problem of variance change point for data on a surface. This change point would suggest a deterioration of the organ to the non-viable status. The statistical challenge here is the development of an efficient procedure that can simultaneously estimate a smooth mean trend on a 2D surface and detect the change point in the variance function on the surface. A naive adoption of the existing methods can result in substantial computational difficulty since the data were collected at a dense grid on a 2D surface. We devise an efficient method that combines subsampling with thin-plate spline smoothing and variance change point detection. Simulations are performed to verify its empirical performance and an application to the organ procurement data is provided.

E0638: **Proportional hazards model with time-dependent covariates measured with error at informative observation times**
Presenter: **Xiao Song**, University of Georgia, United States
The proportional hazards model with time-dependent covariates measured with error at informative observation times under shared random effects models is considered. Although various approaches have been proposed to deal with measurement error for time-dependent covariates, very limited research has been done when the observation times are informative. We propose a new corrected score estimator that allows the observation times to depend on the survival time, the random effects, or other covariates. Compared to existing conditional score and corrected score approach, it relaxes the requirement on non-informative observation times, may substantially improve the efficiency, and is much more robust to deviations from normality of the error. The performance of the estimator is evaluated via simulation studies and by application to data from an HIV clinical trial.

EO172 Room G4701 HIGH-DIMENSIONAL STATISTICS Chair: Emre Barut

E0331: **Practical methods for large and complex data**
Presenter: **Johannes Lederer**, University of Washington, United States
Driven by the advances in technology, large and complex data have become the rule rather than the exception. We introduce novel approaches for the analysis of such data. These approaches apply to a variety of settings, including IV models and network learning. To avoid digression, however, we illustrate the key ideas mainly at the example of feature selection in regression.

E0339: **Influential features PCA for high dimensional clustering**
Presenter: **Wanjie Wang**, National University of Singapore, Singapore
Co-authors: Jiashun Jin, Tracy Ke
Clustering is a major problem in statistics with many applications. In the Big Data era, it faces two main challenges: (1) The number of features is much larger than the sample size; (2) The signals are sparse and weak, masked by large amount of noise. We propose a new tuning-free clustering procedure for large-scale data, Important Features PCA (IF-PCA). IF-PCA consists of a feature selection step, a PCA step, and a k-means step. The first two steps reduce the data dimensions recursively, while the main information is preserved. As a consequence, IF-PCA is fast and accurate, producing competitive performance in application to 10 gene microarray data sets. We also propose a model that can capture the rarity and weakness of signal. Under this model, the statistical limits for the clustering problem and IF-PCA has been found.

E0386: **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.

E0667: **A Neyman-Pearson approach to feature ranking**
Presenter: **Xin Tong**, University of Southern California, United States
Binary classification problems arise frequently in biomedical applications, such as cancer diagnosis using gene expression data. An important question in both basic science research and clinical applications is what genes have the highest predictive power for a certain type of cancer because these genes are possibly cancer driver genes that may serve as treatment targets and/or biomarkers that may improve diagnosis accuracy. Cancer diagnosis belongs to the type of binary classification where the two types of misclassification errors do not have the same priority, because misclassifying a diseased patient as healthy vs. misclassifying a healthy patient as disease would result in severely different consequences. We propose a feature ranking method under the NP paradigm, NP-Rank, motivated by the cancer diagnosis. NP-Rank ranks features based on their type II errors (the less severe type of misclassification error) with their type I errors (the more severe type of error) controlled under a user-specified threshold with high probability. NP-Rank has desirable theoretical guarantees when used with density plug-in classifiers. Extensive numerical studies show that NP-Rank, used with popular classification methods such as Logistic regression, outperforms traditional ranking methods under the classical paradigm.

EO012 Room LT-11 LARGE-SCALE MODELING AND PREDICTION OF FINANCIAL ASSET RETURNS Chair: Marc Paolella

E0502: **Risk parity portfolio allocation under non-Gaussian returns**
Presenter: **Patrick Walker**, University of Zurich, Switzerland
Co-authors: Marc Paolella
A new method for fast computation of large-scale risk parity portfolios under heavy tailed returns is proposed. Asset returns are modeled with an elliptical multivariate generalized hyperbolic distribution, allowing for fast model estimation and a semiclosed-form solution of the risk contributions. Risk is measured by expected shortfall and conditions are given under which this risk parity portfolio coincides with the one when the variance is used. Exploiting several numerical shortcuts, we can compute the risk parity portfolio exceptionally fast and with high precision. An empirical out-of-sample analysis shows that accounting for heavy tails in risk parity allocations leads to improved portfolio performance and lower drawdown. However, the risk parity strategy is dominated by the minimum expected shortfall portfolio in terms of lower risk and higher Sharpe ratio. The portfolio turnover and proportional transaction costs of the competing strategies are investigated. Regularisation of the objective functions is shown to decrease the impact of transaction costs on the net Sharpe ratios of both strategies. The popular equally weighted portfolio is outperformed even under extreme levels of transaction fees. Additionally, we consider using a GARCH-CCC model, instead of the i.i.d. assumption, in order to investigate the impact of heteroskedasticity on the risk parity portfolio, both under Gaussian and heavy tailed returns.

E0624: **A robust knock-out strategy for an high-dimensional portfolio choice problem**
Presenter: **Marco Gambacciani**, University of Zurich and Swiss Financial Institute, Switzerland

A stylized fact of financial returns is that extreme values are part of their historical behavior, while sometimes mistakenly referred to as outlying observations or "outliers". In opposition to other scientific fields, the outliers in financial applications should not be discarded as measurement error, but considered as useful information. The use of robust statistics still provide a very useful and viable tool for application with financial data, as such as optimal portfolio applications, although in a completely different fashion than how is implemented in other fields. We intend to exploit the difference between classical and robust estimators as a provider of useful information about market conditions, which is then converted in proposed trading strategies. With an empirical application to weekly returns of the constituents of the MSCI World Developed Single Stocks Index, we show that the portfolio performance can benefit by incorporating standard portfolio allocation methods with the information derived from comparing robust and classical estimators. In particular, the methods introduced are shown to be viable as knock-out strategies, i.e. starting with a large number of stocks, the strategy is able to do a pre-selection of a group with the most performing stocks, to which the final portfolio allocation method will be applied.

E0734: **Regularized semiparametric estimation of vast dynamic conditional covariance matrices**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

A three-step estimation strategy for dynamic conditional correlation models is proposed. In the first step, conditional variances for individual and aggregate series are estimated by means of QML equation by equation. In the second step, conditional covariances are estimated by means of the polarization identity and consistent estimates of the conditional correlations are obtained by their usual normalization. In the third step, the two-step conditional covariance and correlation matrices are regularized by means of a new non-linear shrinkage procedure and used as starting value for the maximization of the joint likelihood of the model. This yields the final, third step smoothed estimate of the conditional covariance and correlation matrices. Due to its scant computational burden, the proposed strategy allows to estimate vast conditional covariance and correlation matrices. An application to financial data is also provided.

E0800: **COMFORT-Able finance: Extensions of a paradigm for large-scale modeling of asset returns and portfolio construction**
Presenter: **Marc Paolella**, University of Zurich, Switzerland

There are several aspects of financial asset portfolio construction relevant for success. First, the methodology should be applicable to a reasonably large number of assets, at least on the order of 100. Second, calculations should be computationally feasible, straightforward, and fast. Third, realistic transaction costs need to be taken in account for the modeling paradigm to be genuinely applicable. Fourth, and arguably most importantly, the proposed methods should demonstrably outperform benchmark models such as the equally weighted portfolio, Markowitz IID and Markowitz using the DCC-GARCH model. A fifth "icing on the cake" is that the underlying stochastic process assumption is mathematically elegant, statistically coherent, and allows analytic computation of relevant risk measures for both passive and active risk management. The model structure to be shown, referred to as "COMFORT", satisfies all these criteria. Various potential new ideas will also be discussed, with the aim of enticing and motivating other researchers to collaborate and/or improve upon the shown investment vehicles.

EO202 Room LT-12 RECENT ADVANCES IN LARGE-SCALE INFERENCE Chair: Bhaswar Bhattacharya

E0534: **Clustering and feature screening via L1 fusion penalization**
Presenter: **Peter Radchenko**, University of Sydney, Australia
Co-authors: Gourab Mukherjee, Trambak Banerjee

The aim is to study the large sample behavior of a convex clustering framework, which minimizes the sample within cluster sum of squares under an L1 fusion constraint on the cluster centroids. Our analysis is based on a novel representation of the sample clustering procedure as a sequence of cluster splits determined by a sequence of maximization problems. We use this representation to provide a simple and intuitive formulation for the population clustering procedure. We then demonstrate that the sample procedure consistently estimates its population analogue and we derive the corresponding rates of convergence. On the basis of the new perspectives gained from the asymptotic investigation, we propose a key post-processing modification of the original clustering framework. We show, both theoretically and empirically, that the resulting approach can be successfully used to estimate the number of clusters in the population. We also propose an approach for feature screening in the clustering of massive datasets, in which both the number of features and the number of observations can potentially be very large.

E0532: **Bootstrapping spectral statistics in high dimensions**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Andrew Blandino, Alexander Aue

Spectral statistics play a central role in many multivariate testing problems. It is therefore of interest to approximate the distribution of functions of the eigenvalues of sample covariance matrices. Although bootstrap methods are an established approach to approximating the laws of spectral statistics in low-dimensional problems, these methods are relatively unexplored in the high-dimensional setting. The aim is to focus on linear spectral statistics (LSS) as a class of "prototype statistics" for developing a new bootstrap method in the high-dimensional setting. In essence, the method originates from the parametric bootstrap, and is motivated by the notion that, in high dimensions, it is difficult to obtain a non-parametric approximation to the full data-generating distribution. From a practical standpoint, the method is easy to use, and allows the user to circumvent the difficulties of complex asymptotic formulas for LSS. In addition to proving the consistency of the proposed method, we provide encouraging empirical results in a variety of settings. Lastly, and perhaps most interestingly, we show through simulations that the method can be applied successfully to statistics outside the class of LSS, such as the largest sample eigenvalue and others.

E0711: **Nonparametric inference for treatment effects in instrumental variable models**
Presenter: **Bhaswar Bhattacharya**, University of Pennsylvania, United States

In randomized experiments with noncompliance, instrumental variable methods allow for inference about the treatment effect, by controlling for unmeasured confounding. However, many studies do not consider the observed compliance behavior in the testing procedure, which can lead to a loss of power. A novel nonparametric likelihood approach, referred to as the binomial likelihood method, will be discussed. This incorporates the information on compliance behavior while overcoming several limitations of previous techniques and utilizing the advantage of likelihood methods. The proposed method produces proper estimates of the counterfactual distribution functions by maximizing the binomial likelihood over the space of distribution functions. Using this we construct a binomial likelihood ratio test for the null hypothesis of no treatment effect. Asymptotic expansions of the test statistic are derived, and its finite-sample performance is illustrated in simulations and on real data sets.

E0715: **Confounder adjustment in large-scale linear structural models**
Presenter: **Qingyuan Zhao**, University of Pennsylvania, United States
Co-authors: Yang Song

Consider large-scale studies in which thousands of parameters need to be estimated or tests need to be performed simultaneously. In some of these studies, the usual linear regression can be severely biased by latent confounding factors. Two applied examples will be considered. The first is multiple hypothesis testing in genomic screening, where the confounding factors might include batch effect and unmeasured environmental

variables. The second example is evaluation of the performance of mutual funds, where the confounders are any systematic risk factors that are not included in a standard factor model (such as the Fama-French-Carhart four factor model). A two-step procedure based on factor analysis and robust regression will be proposed, and some theoretical guarantees will be given. The statistical method will be applied to a mutual fund return dataset.

EO081 Room LT-13 ADVANCES IN FINANCIAL TIME SERIES ANALYSIS Chair: Mike So

E0422: **Risk premia dynamics of the Japanese financial markets**
Presenter: **Masato Ubukata**, Meiji Gakuin University, Japan
Co-authors: Torben Andersen, Viktor Todorov
The focus is on the predictability of the aggregate stock market returns in Japan. The Japanese market is notoriously difficult to forecast using standard predictive indicators, that are successful for other national indices. Specifically, we test whether the diffusive and jump components of the variance risk premium predict the Japanese returns. Using largely nonparametric risk measures for three major indices, S&P 500, Nikkei 225 and FTSE 100, we first show that country-specific regressions for Japan – contrary to the other countries – produce insignificant predictability patterns. Second, however, we also show that the US left jump variation (LJV) – a proxy for the fear component of the tail risk premium – helps forecast the Japanese excess returns, especially when measured in US dollars. Thus, the dollar-denominated Japanese returns are predictable through the identical mechanism as for the other indices. Moreover, there is a large degree of foreign ownership of Japanese equities. This suggests that the Japanese market is well integrated with the global markets, and is priced accordingly. Third, consistent with the reasoning above, we also find that the US LJV, or the discrepancy between the US and Japanese LJV, have explanatory power for the dollar-yen exchange rate returns.

E0509: **Bayesian modelling and forecasting of value-at-risk via threshold realized volatility**
Presenter: **Cathy W-S Chen**, Feng Chia University, Taiwan
Co-authors: Toshiaki Watanabe
A threshold realized GARCH is proposed that jointly models daily returns and realized volatility, thereby taking into account the bias and asymmetry of realized volatility. We incorporate this threshold realized GARCH model with skew Student-t innovations as the observation equation, view this model as a sharp transition model, and treat the realized volatility as a proxy for volatility under this nonlinear structure. Through the Bayesian Markov chain Monte Carlo method, the model can jointly estimate the parameters in the return equation, the volatility equation, and the measurement equation. As an illustration, we conduct a simulation study and apply the proposed method to the U.S. and Japan stock markets. Based on quantile forecasting and volatility estimation, we find that the threshold heteroskedastic framework with realized volatility successfully models the asymmetric dynamic structure. We also investigate the predictive ability of volatility by comparing the proposed model with traditional GARCH as well as some popular asymmetric GARCH and realized GARCH models. This threshold realized GARCH model with skew Student-t innovations outperforms the competing risk models in out-of-sample volatility and VaR forecasting.

E0542: **Estimation for affine term structure with smooth transition**
Presenter: **Shingo Mukunoki**, Osaka University, Japan
Co-authors: Kosuke Oya
Affine dynamic term structure (ADTS) models suffer difficulty of estimations due to highly non-linear and badly behaved objective function. Although ADTS provides closed-form solutions for yields and bond prices for any maturity, it is too simple to capture the risk sensitivities of market participants. To gauge the risk sensitivity, we consider the ADTS model with time-varying market price of risk. In our model specification, risk factors follow VAR with time-varying coefficients and the market price of risk is represented as the affine form whose coefficients vary over time. Estimation method for our model is based on the asymptotic least squares (ALS) which incorporates the no-arbitrage conditions we must impose to conquer the computational difficulties. We apply the ADTS model with time-varying market price of risk to Japanese government bonds. We find that there are two factors interpreted as level and curvature to explain the term structure of interest rates and their market price of risk are sensitive to the monetary policy. We confirm that the proposed model captures the market risk sensitivity.

E0587: **High-dimensional dynamic covariance modeling via risk factors mapping**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong
The aim is to explore a modified method of dynamic covariance estimation via risk factors mapping. One important feature of the method is to be able to handle dependence estimation of assets of a large portfolio with high computational efficiency. The main idea is to apply a multivariate generalized autoregressive conditional heteroscedasticity (MGARCH) model to a small number of risk factors, which explain the movement of portfolio returns. The idea of risk mapping is demonstrated by an empirical study with a focus on the Hong Kong stock market. Assessment using portfolio risk calculation is also discussed.

EO016 Room LT-14 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Ci-Ren Jiang

E0189: **A simple method to construct confidence bands in functional linear regression**
Presenter: **Masaaki Imaizumi**, Institute of Statistical Mathematics, Japan
Co-authors: Kengo Kato
A simple method is developed 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 first time that confidence bands having theoretical justifications for the PCA-based estimator are derived. 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.

E0237: **Mixture inner product spaces and their application to functional data analysis**
Presenter: **Zhenhua Lin**, University of California, Davis, United States
Co-authors: Hans-Georg Mueller, Fang Yao
The aim is to introduce the concept of mixture inner product spaces associated with a given separable Hilbert space, which feature an infinite-dimensional mixture of finite-dimensional vector spaces and are dense in the underlying Hilbert space. For functional data, mixture inner product spaces provide a new perspective, where each realization of the underlying stochastic process falls into one of the component spaces and is represented by a finite number of basis functions, the number of which corresponds to the dimension of the component space. In the mixture representation of functional data, the number of included mixture components used to represent a given random element is specifically adapted to each random trajectory and may be arbitrarily large. Key benefits of this novel approach are, first, that it provides a new perspective on the construction of a probability density in function space under mild regularity conditions, and second, that individual trajectories possess a trajectory-specific dimension that corresponds to a latent random variable, making it possible to use a larger number of components for less smooth and

a smaller number for smoother trajectories. This enables flexible and parsimonious modeling of heterogeneous trajectory shapes. We establish estimation consistency of the functional mixture density and introduce an algorithm for fitting the functional mixture model based on a modified expectation-maximization algorithm.

E0530: **Correcting selection bias via functional empirical Bayes**
Presenter: **Yingying Fan**, University of Southern California, United States
Consider the problem of estimating mean curves in the setting of functional data, where each observed functional curve can be decomposed into a random mean curve and a error curve around the mean curve. Instead of using the naive method which simply estimates the mean curves using the observed curves, we propose a new method, Functional Empirical Bayes, to reduce the estimation bias. We theoretically study the proposed method and use real data sets to demonstrate the performance of it.

E0785: **Covariance estimation and principal component analysis for spatially dependent functional data**
Presenter: **Yehua Li**, University of California at Riverside, United States
Spatially dependent functional data collected under a geostatistics setting are considered, where locations are sampled from a spatial point process and a random function is observed at each location. Observations on each function are made on discrete time points and contaminated with measurement errors. The error process at each location is modeled as a non-stationary temporal process rather than white noise. Under the assumption of spatial isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrow information from neighboring functions. Under a unified framework for both sparse and dense functional data, where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. The proposed methods are illustrated by simulation studies and a motivating example of the home price-rent ratio data in the New York metropolitan area.

EO257 Room LT-16 BAYESIAN METHODS: NOVEL APPLICATIONS Chair: Michele Guindani

E0547: **A hierarchical nonparametric approach for robust graphical modelling**
Presenter: **Raffaele Argiento**, University of Torino, Italy
Useful tools to express multivariate network structures in gene expression studies are graphical models. However, alternative models are needed when data are strongly overdisperse. An interesting proposal has been previously introduced which uses a Dirichlet process to cluster data-components and accommodate for overdispersion. 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 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. We will compare the performances of the proposed model with competitors via a simulation study, moreover we will explore genomic expression patterns in the yeast Saccharomyces cerevisiae responding to diverse environmental transitions. We will identify the multivariate network structure of the data and meanwhile cluster components according their degree of over dispersion.

E0445: **Probabilistic analysis of multi-way random functions**
Presenter: **Donatello Telesca**, UCLA, United States
Experimental and observational settings are considered, where data can be conceptualized as independent realizations of a univariate or multivariate functional process defined over multiple arguments. We discuss how the probabilistic formulation of latent random features can be useful in characterizing several classes of covariance operators, which, in turn, describe the second order properties of the underlying stochastic processes. In this setting, we show how regularized estimation can be achieved naturally within the Bayesian inferential framework. Several applications to longitudinal functional data and multi-way imaging data are used for illustration.

E0650: **Clustering longitudinal biomarker data using Dirichlet process mixtures**
Presenter: **Wesley Johnson**, UC Irvine, United States
The aim is to jointly model longitudinal diagnostic outcome data (ELISA (continuous) and Fecal Culture (dichotomous)) for individuals that are not infected at the beginning of time, but where some individuals become infected over the course of the study. Individuals that remain uninfected throughout the study have continuous serologic responses that continue to vary about their own baseline level. Responses after infection ultimately increase and then tend to plateau at a higher baseline level. Infection times are modeled using change points, and curve shapes for infected individuals (after unknown time of infection) are modeled using a Dirichlet process mixture of 5 parameter sigmoid curves. This permits clustering of curve shapes, thus allowing for some individuals to a more rapid increase and or shape in response curve than others. All of this is helpful for estimating receiver operating characteristic curves for use of the biomarker in determining cutoffs for detection of infection for individuals in different groups, and for understanding the behavior of testing procedures in different populations where individuals may be known to have been infected longer or shorter periods of time.

E0717: **Network meta-analysis for adverse events: A discrete multivariate Bayesian approach with Gaussian copulas**
Presenter: **Rebecca Graziani**, Bocconi University, Italy
Co-authors: Sergio Venturini
A Bayesian multivariate network meta-analysis (NMA) model of multiple discrete correlated outcomes is proposed. An NMA makes it possible to combine all the direct evidence with all the indirect evidence coming from the studies included in the analysis. While the literature on univariate NMA is now extensive, few methods have been published for synthesizing evidence from studies reporting on multiple discrete outcomes for networks of competing treatments. We propose a new Bayesian copula-based method for multivariate NMA of multiple discrete correlated outcomes. The observed outcome in each study is assumed to be a realization of a multivariate discrete random variable whose elements are marginally distributed according to a binomial distribution. The dependence among the univariate outcomes is induced through a Gaussian copula. The probability to observe any of the individual outcome in each study is modeled as a logistic regression with study-specific baseline effects and arm-specific treatment effects. Estimation proceeds by Markov chain Monte Carlo methods using a mixed Gibbs and adaptive random walk Metropolis-Hastings update for the parameters. The correlation matrix of the Gaussian copula is instead updated through a two-stage parameter expanded Metropolis-Hastings algorithm. We compare the performance of our method with those of other published methods within a simulation study. We apply our proposal to a real data set of adverse events.

EO006 Room LT-17 NON-CAUSAL TIME SERIES MODELS Chair: Alain Hecq

E0190: **Predictive distribution of anticipative alpha-stable Markov processes**
Presenter: **Sebastien Fries**, Hadamard PhD School in Mathematics (Paris-Saclay) and Crest, France
The anticipative, or noncausal, alpha-stable autoregression of order 1 (AR(1)) is a stationary Markov process undergoing short-lived explosive episodes akin to bubbles in financial time series data: recurrently, it diverges away from central values at exponential speed and brutally collapses. Although featuring infinite variance, conditional moments up to integer order four may exist. Little is known about their forms and this impedes understanding of the dynamics of anticipative processes and the ability to forecast them. We provide the functional forms of the conditional expectation, variance, skewness and kurtosis at any forecast horizon under any admissible parameterisation of the process. During bubble episodes, the moments become equivalent to that of a weighted Bernoulli distribution charging complementary probabilities to two polarly-opposite deterministic paths: pursued explosion or collapse. These results extend to the continuous time analogue of the AR(1), the anticipative alpha-stable Ornstein-Uhlenbeck. The proofs build heavily on and extend properties of arbitrary, skewed alpha-stable bivariate random vectors. Other moving averages are considered such as the anticipative AR(2) and the aggregation of anticipative AR(1).

E0235: **Simulation, estimation and selection of mixed causal-noncausal autoregressive models: The MARX package**
Presenter: **Sean Telg**, Maastricht University, Netherlands
Co-authors: Alain Hecq, Lenard Lieb, Sean Telg
The MARX package is presented for the analysis of mixed causal-noncausal autoregressive processes with possibly exogenous regressors. The distinctive feature of MARX models is that they abandon the Gaussianity assumption on the error term. This deviation from the Box-Jenkins approach allows researchers to distinguish backward- (causal) and forward-looking (noncausal) stationary behavior in time series. The MARX package offers functions to simulate, estimate and select mixed causal-noncausal autoregressive models, possibly including exogenous regressors.

E0243: **Detecting time reversibility using quantile autoregressions**
Presenter: **Li Sun**, Maastricht University, Netherlands
Co-authors: Alain Hecq
The aim is twofold. First we propose to detect time irreversibility in stationary time series using quantile autoregressive models (QAR). This approach provides an alternative way to look at the identification of causal from noncausal models. Although we obviously assume that non-Gaussian disturbances generate series we do not need any parametric distribution to maximize (e.g. the Student or the Cauchy) likelihood. This is very interesting for skewed distributions for instance. Secondly, we propose to extend QAR models to QMAR, namely quantile regressions in reverse time. This new modelling is appealing for investigating the presence of bubbles in economic and financial time series. We illustrate our analysis using hyperinflation episodes in Latin American countries.

E0250: **Bootstrap inference under random distributional limits**
Presenter: **Giuseppe Cavaliere**, University of Bologna, Italy
Co-authors: Iliyan Georgiev
Asymptotic bootstrap validity is usually understood as consistency of the distribution of a bootstrap statistic, conditional on the data, for the unconditional limit distribution of a statistic of interest. From this perspective, randomness of the limit bootstrap measure is regarded as a failure of the bootstrap. Nevertheless, apart from an unconditional limit distribution, a statistic of interest may possess a host of (random) conditional limit distributions. This allows the understanding of bootstrap validity to be widened, while maintaining the requirement of asymptotic control over the frequency of correct inferences. First, we provide conditions for the bootstrap to be asymptotically valid as a tool for conditional inference, in cases where a bootstrap distribution estimates consistently, in a sense weaker than the standard weak convergence in probability, a conditional limit distribution of a statistic. Second, we prove asymptotic bootstrap validity in a more basic, on-average sense, in cases where the unconditional limit distribution of a statistic can be obtained by averaging a (random) limiting bootstrap distribution. As an application, we establish rigorously the validity of fixed-regressor bootstrap tests of parameter constancy in linear regression models.

EO079 Room LT-18 STATISTICAL MODELING AND INFERENCE FOR STOCHASTIC PROCESSES Chair: Kengo Kamatani

E0474: **Testing the absence of lead-lag effects in high-frequency data**
Presenter: **Yuta Koike**, University of Tokyo, Japan
The focus is on the problem of testing whether there exists a (possibly) time-lagged correlation between two Brownian motions based on their high-frequency observation data where the observation times are possibly non-synchronous. The test statistic considered here is the maximum of the absolute value of the empirical cross-covariance function as a contrast function to estimate the time-lag parameter of the lead-lag relationship considered here. The approximation of the null distribution of the test statistic is analytically difficult, so we develop a bootstrap procedure to solve this issue. The validity of the proposed bootstrap procedure is ensured by a version of the Gaussian approximation theory.

E0518: **Data driven time scale for ergodic diffusion processes in YUIMA package**
Presenter: **Shoichi Eguchi**, Osaka University, Japan
The aim is to show a parametric estimation of ergodic diffusion processes with unknown sampling stepsize and how to construct estimators of model parameters and sampling stepsize in a fully explicit way. Based on this, we create the function which can estimate model parameter and sampling stepsize for ergodic diffusion processes in R package yuima. We will first overview the estimation method of model parameters and sampling stepsize and then explain the specification of the created function. Some numerical examples are given in order to show how to use the function.

E0533: **Estimation of jump diffusion models by Jarque-Bera normality test**
Presenter: **Yuma Uehara**, The Institute of Statistical Mathematics, Japan
Co-authors: Hiroki Masuda
The estimation problem of jump diffusion models based on high-frequency samples is considered. A well-known estimation approach for the models is to use a jump detection threshold. However, in practical use, the choice of the threshold still remains as an annoying problem. To avoid such a difficulty, we propose a new estimation method based on Jarque-Bera normality test. We will present some numerical experiments and the theoretical properties of our method.

E0549: **Bayesian inference for Stable Lévy driven stochastic differential equations with high-frequency data**
Presenter: **Kengo Kamatani**, Osaka University, Japan
Co-authors: Hiroki Masuda, Ajay Jasra
The focus is on parametric Bayesian inference for stochastic differential equations (SDE) driven by a pure-jump stable Lévy process, which is observed at high frequency. In most cases of practical interest, the likelihood function is not available, so we use a quasi-likelihood and place an associated prior on the unknown parameters. It is shown under regularity conditions that there is a Bernstein-von Mises theorem associated to the

posterior. We then develop a Markov chain Monte Carlo (MCMC) algorithm for Bayesian inference and assisted by our theoretical results, we show how to scale Metropolis-Hastings proposals when the frequency of the data grows, in order to prevent the acceptance ratio going to zero in the large data limit. Our algorithm is presented on numerical examples that help to verify our theoretical findings.

EO261 Room P4701 LATENT VARIABLE MODELS AND PSYCHOMETRICS Chair: Gongjun Xu

E0173: **Conditional dependence among items in DINA model: Application of the multivariate probit model**
Presenter: **Kevin Carl Santos**, The University of Hong Kong, Hong Kong
Co-authors: Alexander de Leon, Jimmy de la Torre, Mingchen Ren
The deterministic input, noisy "and" gate (DINA) model is a tractable and interpretable cognitive diagnosis model that can be used to identify students' mastery and nonmastery of skills in a subject domain of interest. We introduce the multivariate probit DINA (DINA-MP) model, which is a re-formulation of the DINA model that can account for potential relationships between test items that may remain even after conditioning on the students' underlying skills. A computationally efficient parameter-expanded Monte Carlo EM (PX-MCEM) algorithm is outlined for maximum likelihood estimation of the parameters of the proposed model. A simulation study is conducted to determine the extent to which ignoring between-item conditional dependence may yield biased estimates with understated standard errors, thus yielding confidence intervals that are misleadingly narrow and tests with inflated Type I errors. Finally, fraction subtraction data are analyzed to examine the practical viability of the DINA-MP model and the corresponding PX-MCEM algorithm.

E0212: **Item calibration methods for multi-stage design**
Presenter: **Chun Wang**, University of Minnesota, United States
Many large scale educational surveys have moved from linear form design to multistage testing (MST) design. A MST tailors the set of items (i.e., target block) a student sees to the students individual ability level, so that no examinee receives too many overly easy or difficult items. Consequently, MST can provide more accurate latent trait estimates (θ) using fewer items than required by linear tests. However, MST generates incomplete response data by design; hence questions remain as to whether the item calibration procedure for the traditional linear form can still apply? If not, what adjustments should be made? Deriving from the missing data mechanism, two new item calibration methods will be proposed, namely the multiple-group with truncated normal prior and the multiple-group with empirical histogram prior. They will be compared to the traditional single-group marginal maximum likelihood (MMLE) and multiple-group MMLE in terms of item parameter recovery.

E0299: **A fused latent and graphical model**
Presenter: **Jingchen Liu**, Columbia University, United States
One of the main tasks of statistical models is to characterize the dependence structures of multi-dimensional distributions. Latent variable model takes advantage of the fact that the dependence of a high dimensional random vector is often induced by just a few latent (unobserved) factors. Such models are employed in the analysis of marketing, e-commerce, social network, and many other fields where human behaviors are observed and are summarized to a few characteristics. We present several examples. In these examples, a common problem is that the dimension grows higher and the dependence structure becomes more complicated. It is hardly possible to find a low dimensional parametric latent variable model that fits well. We enrich the model by including a graphical structure on top of the latent structure. The graph captures the remaining dependence and is often more interpretable than graphs built on marginal dependence.

E0426: **A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests**
Presenter: **Chia-Yi Chiu**, Rutgers University, United States
The Q-matrix of a cognitively diagnostic test is said to be complete if it allows for the identification of all possible proficiency classes among examinees. Completeness of the Q-matrix is therefore a key requirement for any cognitively diagnostic test. However, completeness of the Q-matrix is often difficult to establish, especially, for tests with a large number of items involving multiple attributes. As an additional complication, completeness is not an intrinsic property of the Q-matrix, but can only be assessed in reference to a specific cognitive diagnosis model (CDM) supposed to underly the data; that is, the Q-matrix of a given test can be complete for one model but incomplete for another. A method is presented for assessing whether a given Q-matrix is complete for a given CDM. The proposed procedure relies on the theoretical framework of general CDMs and is therefore legitimate for CDMs that can be reparameterized as a general CDM.

EO051 Room P4703 RECENT ADVANCES IN MODELLING AND CLUSTERING VIA MIXTURE MODELS Chair: Geoffrey McLachlan

E0411: **A mixture regression model of multivariate generalized Bernoulli distributions**
Presenter: **Shu-Kay Ng**, Griffith University, Australia
In healthcare research, outcome variables in a categorical form are commonplace. Data collection often involves the acquisition of information on a spectrum of individuals feature variables (risk factors) that may influence the outcomes. Clustering of individuals based on the categorical outcome variables (p dimensions) and associated vector of q-dimensional risk factors can be obtained via a mixture regression model-based approach. With this approach, each component-density function is specified by a multivariate generalized Bernoulli distribution consisting of one draw on d_i categories for each outcome variable $i = 1, \ldots, p$. Moreover, the risk factors are included in the mixing proportion via a logistic model. The proposed mixture regression model is thus able to simultaneously cluster individuals into groups with different patterns of outcomes and identify the characteristics of individuals that are relevant for explaining the heterogeneity in outcome patterns. Parameter estimation is based on maximum likelihood via the expectation-maximization (EM) algorithm. This model can also be adopted to cluster mixed categorical and continuous data, and apply in consensus clustering where each categorical outcome variable represents a partition of individuals based on a number of different sets of feature variables. The method is illustrated using simulated data and a publicly available data set concerning comorbidity patterns among alcohol- and drug-dependent adults.

E0527: **Simultaneous detection of differential gene expression and gene clustering using mixture models**
Presenter: **Andrew Jones**, University of Queensland, Australia
Co-authors: Geoffrey McLachlan
The purpose is to examine the use of EMMIXcontrasts, a finite mixture model based approach for the detection of differential gene expression, which simultaneously clusters the genes based on their expression profiles. This method has been previously been applied to microarray expression data and has now also been shown to work on suitably transformed RNA-sequence data, further enhancing the method's applicability. This method is demonstrated on a number of datasets, via the revamped EMMIXcontrasts R Package, for both the detection of differential gene expression and gene clustering.

E0594: **Assessment of model fit in clustering via mixture models**
Presenter: **Suren Rathnayake**, The University of Queensland, Australia
Some issues are considered associated with the fitting of a normal mixture model to cluster data known to be drawn from an unknown number of distinct classes. The use of the likelihood ratio statistic is investigated for tests on the smallest number of components in the mixture model for it

to be compatible with the observed data. Also, under the implicit assumption that the clusters implied by the fitted mixture model are in correct correspondence with the external existing classes, we investigate further the estimation of the accuracy of the implied clustering. For this purpose, an estimator of the overall correct allocation rate is formed by averaging the maximum of the (estimated) posterior probabilities of component membership for each observation.

E0682: **Mixture modelling with scale mixtures of skew normal distributions**
Presenter: **Sharon Lee**, University of Queensland, Australia
In recent years, mixture models with skew component distributions have received increasing attention. The literature now offers a wide variety of non-normal distributions with different properties suitable for a range of applications. An overview of existing skew models is provided, focusing on those adopted in the model-based clustering literature. We then consider a very general family of skew distributions, namely, the scale mixture of canonical fundamental skew normal (SMCFUSN) distributions. This family encapsulates many important and commonly used symmetric and skew distributions including the normal, t, hyperbolic, slash, and their skew variants. Mixtures of SMCFUSN distributions can be fitted by maximum likelihood via an EM-type algorithm. Dimension reduction via a factor version of mixtures of SMCFUSN distributions is also considered. The usefulness of the approach will be demonstrated via clustering applications to some real datasets.

EO304 Room P4704 RECENT ADVANCE IN (SEMI)PARAMETRIC MODELLING Chair: Thomas Fung

E0457: **Nonparametric tilted function estimation: Some recent development**
Presenter: **Hassan Doosti**, Macquarie University, Australia
The history of nonparametric curve estimation by tilting and its applications is discussed. The target will be the estimation of density functions and regression functions. Then, we will review some recent developments in the field. We will also show that the proposed tilted estimator provides a convergence rate which is strictly faster than the usual rate. The performance of the proposed tilted estimator through both theoretical and numerical studies will be investigated. Finally we propose new applications/open problems of tilting in some applied fields.

E0451: **A weighted partial likelihood approach for zero-truncated models**
Presenter: **Jakub Stoklosa**, University of New South Wales, Australia
Co-authors: Wen-Han Hwang
Motivated by the Rao-Blackwell theorem, we develop a weighted partial likelihood approach to estimate model parameters for the zero-truncated binomial distribution. The resulting estimating function is equivalent to a weighted score function for a standard binomial model, hence allowing for straightforward implementation for estimating model parameters. We evaluate the efficiency for this new approach and show that it performs almost as well as the maximum zero-truncated likelihood method. In addition, the weighted partial likelihood approach can also be extended to zero-truncated Poisson models. An application to estimating population sizes using capture-recapture models is also addressed. This novel approach is then implemented with a corrected score method to accommodate models with measurement error, which has yet to be developed for zero-truncated regression measurement error models. We examine the performance of the proposed methods through simulation studies and real data.

E0589: **Semiparametric modelling in generalized linear models**
Presenter: **Busayasachee Puang-Ngern**, Macquarie University, Australia
Co-authors: Jun Ma, Ayse Bilgin, Timothy Kyng
The semiparametric generalized linear models (SP-GLMs) are an extension of the well-known generalized linear models (GLMs). These semiparametric models are an alternative form of regression modelling. The additional nonparametric components in the conditional response density allow the distribution to be specified by the data while the response distribution is still in the exponential family. Iterative methods are applied to estimate the regression coefficient parameters and the nonparametric components simultaneously. We make a comparison of the biases, the asymptotic standard error, the Monte Carlo standard error, type I error and the likelihood ratio test of the regression coefficients between the SP-GLM and the GLM. Using simulation and a hypothetical parametric GLM to generate a sample of data, we fitted a SP-GLM to that data and found that the SP-GLM provides very similar results to the parametric GLM. Identical results can be found for logistic regression. The SP-GLM can provide more reasonable statistical inference for zero-inflated data.

E0747: **Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood**
Presenter: **Kenny Xu**, Duke-NUS, Singapore, Singapore
Cox proportional hazard models estimation is considered under informative right censored data using maximum penalized likelihood, where dependence between censoring and event times are modelled by a copula function and a roughness penalty function is used to restrain the baseline hazard as a smooth function. Since the baseline hazard is non-negative, we propose a special algorithm where each iteration involves updating regression coefficients by the Newton algorithm and baseline hazard by the multiplicative iterative algorithm. The asymptotic properties for both regression coefficients and baseline hazard estimates are developed. The simulation study investigates the performance of our method and also compares it with an existing maximum likelihood method. We apply the proposed method to a dementia patients dataset.

Tuesday 19.06.2018 16:10 - 17:25 Parallel Session D – EcoSta2018

EO121 Room G4302 FORECASTING/FORECAST COMBINATION Chair: Andrey Vasnev

E0330: **To pool or not to pool: Looking for a good strategy for parameter estimation and forecasting in panel regressions**
Presenter: **Wendun Wang**, Erasmus University Rotterdam, Netherlands
Co-authors: Xinyu Zhang, Richard Paap
The focus is on estimating the slope parameters in potentially heterogeneous panel data regressions for explaining and forecasting cross-country sovereign credit risk. We propose a novel optimal pooling averaging estimator that makes an explicit trade-off between efficiency gains from pooling and bias due to heterogeneity. By theoretically and numerically comparing various estimators, we find that a uniformly best estimator does not exist and that our new estimator is superior in non-extreme cases. The results provide practical guidance for the best estimator depending on features of data and models.

E0450: **Theory and practice of combining forecasts of higher moments in financial data**
Presenter: **Andrey Vasnev**, University of Sydney, Australia
Co-authors: Laurent Pauwels, Peter Radchenko
The aim is to investigate the theory and practice of two novel approaches to combining forecasts of higher moments, specifically skewness and kurtosis, to predict the moments of implicitly and explicitly combined distributions in financial data. Alternative linear and nonlinear models can be combined to forecast the higher moments using explicitly and implicitly combined models. The first approach combines the models using optimal weights to predict the higher moments of an explicitly combined model. The second approach combines the higher moments of the models using optimal weights to predict the higher moments of an implicitly combined model. We expand the knowledge and techniques of combining forecasts by doubling the number of moments that are considered in the literature from two moments, namely the mean and variance, to four, with the addition of the skewness and kurtosis parameters. Several new versions of the Kullback-Leibler (KL) Information Criterion (IC), or KLIC, also known as the KL divergence or discrimination information, are used to derive the optimal weights for combining forecasts of skewness and kurtosis, and hence to evaluate forecast accuracy of alternative moments. Novel variations and extensions of alternative IC, such as the Akaike IC (AIC), which is an estimator of KLIC, and the Schwartz Bayesian IC (SBIC) and Hannan-Quinn Criterion (HQC), neither of which is an estimator of KLIC, will also be analysed.

E0180: **A Mallows-type model averaging estimator for the varying-coefficient partially linear model**
Presenter: **Rong Zhu**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Co-authors: Alan Wan, Xinyu Zhang, Guohua Zou
In the last decade, significant theoretical advances have been made in the area of frequentist model averaging (FMA); however, the majority of this work has emphasised parametric model setups. FMA for the semiparametric varying-coefficient partially linear model (VCPLM) is considered. VCPLM has gained prominence to become an extensively used modeling tool in recent years. Within this context, we develop a Mallows-type criterion for assigning model weights and prove its asymptotic optimality. A simulation study and a real data analysis demonstrate that the FMA estimator that arises from this criterion is vastly preferred to estimators obtained by information criterion score-based model selection and FMA methods. The analysis is complicated by the fact that for the VCPLM, uncertainty is not only with respect to the choice of covariates, but also to the component in the model to which the covariate belongs.

EO119 Room G4701 INFERENCE FOR LARGE COMPLEX DATA Chair: Heng Lian

E0199: **A diagonal likelihood ratio test for equality of mean vectors in high-dimensional data**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong
A likelihood ratio test framework is proposed for testing normal mean vectors in high-dimensional data under two common scenarios: the one-sample test and the two-sample test with equal covariance matrices. We derive the test statistics under the assumption that the covariance matrices follow a diagonal matrix structure. In comparison with the diagonal Hotelling's tests, our proposed test statistics display some interesting characteristics. In particular, they are a summation of the log-transformed squared t-statistics rather than a direct summation of those components. More importantly, to derive the asymptotic normality of our test statistics under the null and local alternative hypotheses, we do not require the assumption that the covariance matrix follows a diagonal matrix structure. As a consequence, our proposed test methods are very flexible and can be widely applied in practice. Finally, simulation studies and a real data analysis are also conducted to demonstrate the advantages of our likelihood ratio test method.

E0280: **A regularized model-based clustering method for image classification**
Presenter: **Ying Zhu**, National Institute of Education, Nanyang Technological University, Singapore
Finite mixture models provide a flexible probabilistic modeling tool to handle heterogeneous data with a finite number of unobserved components. They are employed for model-based clustering in image classification. In high-dimensional spectroscopic data settings, spectral variable selection is both challenging and important to enable the feasibility of multivariate distribution fitting, especially for use in a real-time image classification. A regularized model-based clustering model is presented which enables an automatic selection of a small number of informative spectral variables for image classification. This model is on real life spectroscopic image data. The well-performed selection of spectral features leads to improve the classification accuracy, as well as to substantial reduce the clustering model complexity, and to provide better image representation.

EO113 Room LT-11 DYNAMIC ECONOMETRIC MODELLING Chair: Maria Kyriacou

E0548: **Moment and memory properties of exponential-type conditional heteroscedasticity models**
Presenter: **Xiaoyu Li**, Capital University of Economics and Business, China
Co-authors: James Davidson
The aim is to investigate the moment and memory properties of exponential-type conditional heteroscedasticity models, including exponential generalised autoregressive conditional heteroscedastic (EGARCH) and the fractionally integrated (FIEGARCH(BM)). The moment conditions of these models are derived from previous literature, and the memory properties are measured by using a near-epoch dependence (NED) functions of an independent process approach. The existence of moments supports the limited memory properties of these models. It is shown that the exponential autoregressive conditional heteroscedastic (EARCH)(∞) processes may exhibit geometric memory, hyperbolic memory or long memory. A general expression of the HY/FIEGARCH(DL) model is introduced depending on the properties of the lag coefficients, and the simulation results show that the HYEGARCH model has a hyperbolic memory and the FIEGARCH(DL) model exhibits long memory in the absolute return series. The functional central limit theorem (FCLT) or fractional FCLT for the partial sum of the processes in the EGARCH-type models is also derived.

E0555: **Spatial heterogeneous 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 their inability to capture spatial non-linearities and unobserved heterogeneity. We propose a spatial heterogeneous autoregressive exogenous (SHARX) model captures for such non-linearities and unobserved heterogeneity by allowing for varying-coefficients 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 spatial trends/non-stationarity in y or heterogeneous 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 are 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 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).

E0586: **Modelling nonlinear and fractionally cointegrated price discovery in commodity markets**
Presenter: **Chi Wan Cheang**, University of Southampton, United Kingdom
A fractionally cointegrated vector autoregressive (FCVAR) model with nonlinear disequilibrium adjustment is developed. This property is relevant for the price discovery in commodity markets. In addition to the existence of long memory and normal state of contango or backwardation in the long run spot-futures equilibrium, the commodity spot and futures prices are shown to have nonlinear co-variation dynamics. It occurs that in some time periods the price disequilibrium is more persistence whereas in other periods the disequilibrium is correcting with a fast rate of adjustment. The model is applied to some data used previously to model price discovery in non ferrous metals markets in the UK using the cointegrated VAR model.

EO233 Room LT-13 TOPICS IN FINANCIAL ECONOMETRICS AND FORECASTING Chair: Rachida Ouysse

E0529: **Asset pricing with endogenous state-dependent risk aversion**
Presenter: **Rachida Ouysse**, University of New South Wales, Australia
The evidence on time varying risk aversion parameter in the consumption CAPM using a multinomial Logit model is revisited. We assume that the risk aversion is dependent on the economic conditions through a business cycle dummy variable which depends on a set of latent factors that determine the state of the economy. The model therefore implies two sets of conditional moments: those implied by Euler equation and those from the multinomial equation. We estimate the two systems jointly using the (continuously updated) generalized method of moments (GMM). In this new model, the average risk aversion parameter is determined by the latent variables that define the business cycle. We address issues of identification (or lack of identification) of the multinomial equation and how it translates into the stochastic discount factor.

E0461: **Volatility spillovers and latent network linkages**
Presenter: **Laurent Pauwels**, University of Sydney, Australia
Co-authors: Manabu Asai, Michael McAleer
Volatility spillovers of financial assets is proposed to be modeled using network models. The spillover linkages across assets are modelled with a latent network. The latent network, which informs on the linkages across assets, is identified using Bayesian computational methods. Once the network is identified, the spillover effects across markets can be estimated and statistically tested. This approach reduces the estimation burden of the spillover effects typically encountered in multivariate volatility models with large parameter space.

E0668: **Hierarchical probabilistic forecasting of electricity demand with smart meter data**
Presenter: **Souhaib Ben Taieb**, Monash University, Australia
Co-authors: James Taylor, Rob Hyndman
Electricity smart meters record consumption, on a near real-time basis, at the level of individual commercial and residential properties. From this, a hierarchy can be constructed consisting of time series of demand at the smart meter level, and at various levels of aggregation, such as substations, cities and regions. Forecasts are needed at each level to support the efficient and reliable management of consumption. A limitation of previous research in this area is that it considered only deterministic prediction. To enable improved decision-making, we introduce an algorithm for producing a probability density forecast for each series within a large-scale hierarchy. The resulting forecasts are coherent in the sense that the forecast distribution of each aggregate series is equal to the convolution of the forecast distributions of the corresponding disaggregate series. Our algorithm has the advantage of synthesizing information from different levels in the hierarchy through forecast combination. Distributional assumptions are not required, and dependencies between forecast distributions are imposed through the use of empirical copulas. Scalability to large hierarchies is enabled by decomposing the problem into multiple lower-dimension sub-problems. Results for UK electricity smart meter data show performance gains for our method when compared to benchmarks.

EO204 Room LT-14 COMPUTATION AND INFERENCE WITH LARGE AMOUNTS OF DATA Chair: HaiYing Wang

E0223: **Efficient second-order optimization methods for machine learning**
Presenter: **Fred Roosta**, University of Queensland, Australia
Co-authors: Michael Mahoney
Contrary to the scientific computing community which has, wholeheartedly, embraced the second-order optimization algorithms, the machine learning community has long nurtured a distaste for such methods, in favor of first-order alternatives. We argue that such reluctance to employ curvature information can indeed hinder the training procedure in a variety of ways. Specifically, in the context of non-convex machine learning problems, we demonstrate the theoretical properties as well as empirical performance of a variety of efficient Newton-type algorithms. In the process, we highlight the serious disadvantages of first-order methods and, in their light, showcase the practical advantages offered by such second-order methods.

E0686: **Model robust scenarios for active learning**
Presenter: **Douglas Wiens**, University of Alberta, Canada
Co-authors: Rui Nie, Zhichun Zhai
Experimental design in Statistics is very much like active learning in Machine Learning. In both cases, the idea is that predictor variables are chosen in some optimal manner, and at these values a response variable is observed. In design, the regressors are determined by a design measure, obtained by the designer according to some optimality principle such as minimum mean squared error of the predicted values. In 'passive learning' these regressors are randomly sampled from 'the environment', in active learning they are randomly sampled from a subpopulation according to a probability density derived by the designer in some optimal manner. So a major difference between active learning and experimental design is in the random, rather than deterministic, sampling of the regressors from the learning density or design measure. When the parametric model

being fitted is exactly correct, the corresponding loss functions are asymptotically equivalent and the methods of experimental design apply, with only minor modifications, to active learning. When however this model is in doubt, some significant differences between robust design and robust learning emerge, and with them interesting, new, optimality problems.

E0746: **A Bayesian spatial-temporal model with latent MLG random effects with application to earthquake magnitudes**
Presenter: **Guanyu Hu**, University of Connecticut, United States

A Bayesian spatial-temporal model is introduced for analyzing earthquake magnitudes. Specifically, we define a spatial-temporal Pareto regression model with latent multivariate log-gamma random vectors to analyze earthquake magnitudes. This represents a marked departure from the traditional spatial generalized linear regression model, which uses latent Gaussian random effects. The multivariate log-gamma distribution results in a full-conditional distribution that can be easily sampled from, which leads to a fast mixing Gibbs sampler. Thus, our proposed model is a computationally efficient approach for modeling Pareto spatial data. The empirical results suggest similar estimation properties between the latent Gaussian model and latent multivariate log-gamma model, but our proposed model has stronger predictive properties. Additionally, we analyze a small US earthquake data set as an illustration of the effectiveness of our approach.

EO117 Room LT-15 BAYESIAN HIERARCHICAL MODELS AND COMPUTATIONAL METHODS Chair: Siew Li Linda Tan

E0321: **A divide-and-conquer Bayesian approach to large-scale kriging**
Presenter: **Cheng Li**, National University of Singapore, Singapore
Co-authors: Rajarshi Guhaniyogi, Terrance Savitsky, Sanvesh Srivastava

Flexible hierarchical Bayesian modeling of massive data is challenging due to poorly scaling computations in large sample size settings. The motivation comes from spatial process models for analyzing geostatistical data, which typically entail computations that become prohibitive as the number of spatial locations becomes large. We propose a three-step divide-and-conquer strategy within the Bayesian paradigm to achieve massive scalability for any spatial process model. We partition the data into a large number of subsets, apply a readily available spatial process model on every subset in parallel, and optimally combine the posterior distributions estimated on all the subsets into a pseudo posterior distribution that is used for predictive and parametric inference and residual surface interpolation. We call this approach "Distributed Kriging" (DISK). The Bayes risk of estimating the true residual spatial surface using the DISK posterior distribution decays to zero at a nearly optimal rate under mild assumptions. While DISK is a general approach to divide-and-conquer Bayesian nonparametric regression, we focus on its applications in spatial statistics and demonstrate its empirical performance using models based on stationary full-rank and nonstationary low-rank Gaussian process priors. A variety of simulations and a geostatistical analysis of the Pacific Ocean sea surface temperature data validate our theoretical results.

E0360: **Efficiently combining pseudo marginal and particle Gibbs sampling**
Presenter: **David Gunawan**, University of New South Wales, Australia
Co-authors: Christopher K Carter, Robert Kohn

Particle Markov Chain Monte Carlo methods are used to carry out inference in non-linear and non-Gaussian state space models, where the posterior density of the states is approximated using particles. The correlated pseudo marginal sampler has been recently introduced and it has been shown that it can be much more efficient than the standard pseudo marginal approach. A particle MCMC sampler has also been proposed which generates parameters that are highly correlated with the states using a pseudo marginal method that integrates out the states, while all other parameters are generated using particle Gibbs. We show how to combine these two approaches to particle MCMC to obtain a flexible sampler with a superior performance to each of these two approaches. We illustrate the new sampler using a multivariate factor stochastic volatility model with leverage.

E0424: **Using history matching for prior choice**
Presenter: **Xueou Wang**, Singapore University of Technology and Design, Singapore
Co-authors: David Nott, Christopher Drovandi, Kerrie Mengersen, Michael Evans

It can be important in Bayesian analyses of complex models to construct informative prior distributions which reflect knowledge external to the data at hand. Nevertheless, how much prior information an analyst can elicit from an expert will be limited due to constraints of time, cost and other factors. Effective numerical methods are developed for exploring reasonable choices of a prior distribution from a parametric class, when prior information is specified in the form of some limited constraints on prior predictive distributions, and where these prior predictive distributions are analytically intractable. The methods developed may be thought of as a novel application of the ideas of history matching, a technique developed in the literature on assessment of computer models. We illustrate the approach in the context of logistic regression and sparse signal shrinkage prior distributions for high-dimensional linear models.

EO141 Room LT-16 THEORETICAL PERSPECTIVES FOR BAYESIAN NONPARAMETRICS Chair: Yongdai Kim

E0185: **Bayesian sparse linear regression with unknown symmetric error**
Presenter: **Minwoo Chae**, Case Western Reserve University, United States
Co-authors: Lizhen Lin, David Dunson

Bayesian procedures for sparse linear regression are considered when errors have a symmetric but otherwise unknown distribution. The unknown error distribution is endowed with a symmetrized Dirichlet process mixture of Gaussians. For the prior on regression coefficients, a mixture of point masses at zero and continuous distributions is considered. Asymptotic behavior of the posterior is studied with diverging number of predictors. The compatibility and restricted eigenvalue conditions yield the optimal convergence rate of the regression coefficients in l1 and l2 norms, respectively. The convergence rate is adaptive to both the unknown sparsity level and the unknown symmetric error density. In addition, strong model selection consistency and a semi-parametric Bernstein-von Mises theorem are considered under slightly stronger conditions.

E0472: **Uncertainty quantification and computational methods for the spike and slab prior**
Presenter: **Botond Szabo**, Leiden University, Netherlands
Co-authors: Ismael Castillo, Tim van Erven

Spike and slab priors are frequently used in various fields of applications to induce sparsity in high-dimensional models. It is well known that sampling from the corresponding posterior distribution is computationally very demanding and accurate methods with theoretical guarantees break down on small sample and feature sizes. In practice therefore approximation algorithms were suggested based on optimisation methods. These methods have, however, only limited theoretical underpinning. Firstly, we introduce accurate (analytic) computational methods in the context of the Gaussian sequence model, which can handle moderately large dimensional parameters. Secondly, we will investigate how reliable are the Bayesian uncertainty statements using spike-and-slab priors from a frequentist perspective, again in the context of the Gaussian sequence model. We will derive sufficient and (in some sense) necessary condition under which Bayesian credible sets provide reliable confidence statements.

E0553: **What is asymptotically testable and what is not**
Presenter: **Bas Kleijn**, University of Amsterdam, Netherlands

Given a statistical model for i.i.d. data, certain hypotheses can be tested consistently, while others cannot. If one thinks of consistent tests only in terms of converging sequences of test statistics, some immediate, simple conclusions can be drawn. But classical counterexamples demonstrate that the matter is more involved. We address the problem of what characterizes the asymptotic testability of hypotheses for uniform, pointwise and Bayesian tests. Posteriors distinguish measurable hypotheses (prior-almost-surely), but frequentist tests require more. Application of the Le Cam-Schwartz theorem (write U for the associated uniformity) leads to two equivalences: hypotheses are testable with uniform power if and only if they are separated by a uniformity U. Hypotheses are testable in a pointwise sense, if and only if the testing problem can be represented (continuously with respect to U) in a separable metric space. The above is illustrated with a large number of examples.

EO065 Room LT-17 NEW DEVELOPMENTS IN TIME SERIES ECONOMETRICS Chair: Daniel Preve

E0192: **Asymptotic trimming for importance sampling estimators with infinite variance**
Presenter: **Thomas Yang**, Australian National University, Australia

Importance sampling is a popular Monte Carlo method used in a variety of areas in econometrics. When the variance of the importance sampling estimator is infinite, the central limit theorem does not apply, and estimates tend to be volatile even when the simulation size is large. We consider asymptotic trimming in such a setting. Specifically, we propose a bias-corrected tail-trimmed estimator such that it is consistent and has finite variance. We show that the proposed estimator is asymptotically normal, and has good finite-sample properties in a Monte Carlo study.

E0731: **Solving asset pricing models Via nonparametric two-stage penalized B-spline regression**
Presenter: **Liyuan Cui**, City University of Hong Kong, Hong Kong
Co-authors: Yongmiao Hong, yingxing li

A nonparametric 2SLS Penalized B-spline regression method is presented for unknown policy functions in an exchange economy, which allows the true dynamics of state variables to determine asset prices. Unlike current numerical solution methods, this new method does not require imposing auxiliary assumptions on the conditional distributions or matching the mean and variance of state variables, which enables real state dynamics to determine equilibrium equity prices. We propose a fast generalized cross-validation method to determine the optimal penalization in the 2SLS B-splines regressions, which ensures efficient and consistent estimation of the policy function for a broad class of stationary Markov state variables. The newly proposed regression method will become a pivotal approach for obtaining a consistent estimation of the price-dividend ratio function and equity prices in the presence of misspecified state variables or those with unknown dynamics.

E0591: **A mixture autoregressive model based on Student's t-distribution**
Presenter: **Daniel Preve**, City University of Hong Kong, Hong Kong
Co-authors: Mika Meitz, Pentti Saikkonen

A new mixture autoregressive model based on Student's t-distribution is proposed. A key feature of our model is that the conditional t-distributions of the component models are based on autoregressions that have multivariate t-distributions as their (low-dimensional) stationary distributions. That autoregressions with such stationary distributions exist is not immediate. Our formulation implies that the conditional mean of each component model is a linear function of past observations and the conditional variance is also time varying. Compared to previous mixture autoregressive models our model may therefore be useful in applications where the data exhibits rather strong conditional heteroskedasticity. Our formulation also has the theoretical advantage that conditions for stationarity and ergodicity are always met and these properties are much more straightforward to establish than is common in nonlinear autoregressive models. An empirical example employing a realized kernel series based on S&P 500 data shows that the proposed model performs well in volatility forecasting.

EO030 Room LT-18 STATISTICAL METHODS FOR FUNCTIONAL DATA Chair: Jeng-Min Chiou

E0658: **Estimation and inference of time-varying coefficients in non-linear ordinary differential equation models**
Presenter: **Naisyin Wang**, Univ of Michigan, United States

The use of ordinary differential equations (ODEs) in modeling dynamic systems has gained high popularity in the recent decade. The physiological meanings of ODE parameters are often useful in enabling scientists to gain better understanding of the underlying system. On this regard, both estimation and inference procedures are essential. Even though various time-varying coefficient ODE model has been considered previously. The inference procedures considered earlier tends to be similar to what has been used in parametric models. We propose a new set of estimation and inference procedures for time-varying ODE coefficients. Our methods take into account features that are unique for ODE estimation and, as such, are adaptive in nature. The validity of the proposed procedures is justified through asymptotic properties. The numerical efficacy of the methodologies is illustrated using both synthetic and real-world data-sets.

E0619: **Subspace clustering for functional data**
Presenter: **Yoshikazu Terada**, Osaka University; RIKEN, Japan
Co-authors: Michio Yamamoto

Functional data have the intrinsic high dimensional nature. This nature often makes possible the very good performance in supervised classification for functional data. In the supervised classification problems, it is known that, using the projection into the finite-dimensional subspace, we can extract the intrinsic high dimensional nature from functional data. In the context of unsupervised classification, there are several clustering methods based on the projection into the subspace. However, since these methods mainly focus on within-cluster variance or dimension reduction, the projected data do not necessarily reflect the hidden true cluster structure. A new subspace clustering method for functional data is proposed, which is based on a novel cluster-separation criterion in the finite-dimensional subspace. The proposed method works well not only for the simulated data, but also for the real data which are difficult to obtain a good classification performance by the existing methods.

E0437: **Spatially constrained functional clustering using nearest neighbors**
Presenter: **Yu-Ting Chen**, National Chenchi University, Taiwan
Co-authors: John Aston, Jeng-Min Chiou

Brain architecture is well known to not be simply located randomly and spatially from one brain to another, and spatial information should not be ignored in the analysis of data. Thus, simply determining clusters from their functional responses is insufficient; the spatial location from which they originate should also be considered. We proposed spatially constrained functional data clustering, motivated by the demand for it in electroencephalography (EEG) analysis based on multilevel functional principal component analysis. The objective was to cluster the EEG channels that had similar response patterns in the recorded event-related potentials.

EO231 Room P4302 STOCHASTIC FRONTIER ANALYSIS, HETEROGENEITY AND DEPENDENCE Chair: Artem Prokhorov

E0253: **Firm heterogeneity and dynamic panel stochastic frontier Models**
Presenter: **Hung-pin Lai**, National Chung Cheng University, Taiwan

The estimation of dynamic panel stochastic frontier (DSF) models with firm heterogeneity is considered. We assume that the technical inefficiency follows an AR(1) process, which allows the firm to improve its efficiency from past experiences. Moreover, the model contains two sources of unobserved firm heterogeneity. One is from the heterogeneous distribution of the inefficiency. The other is either from the firm fixed effects, which implies heterogeneity in the production technology; or from the heterogeneity in the speed of adjustment of inefficiency. In the latter case, the AR(1) coefficient is firm specific. We discuss using the likelihood-based approaches to estimate the models. The finite sample performance of the proposed estimators is also investigated by Monte Carlo experiments.

E0788: **A new approach to estimate intergenerational mobility elasticities**
Presenter: **Le Wang**, University of Oklahoma, United States

A nonparametric framework is provided for estimating intergenerational mobility elasticities (IGE) of children's income with respect to parental income. We allow the IGEs to be heterogeneous, by leaving the relationship of parental and child incomes unspecified, while acknowledging and addressing the latent nature of both child and parental permanent incomes and the resulted life-cycle bias. Our framework is a general nonparametric models that can accommodate non-classical measurement errors. Our framework enables us to propose a formal test of the widely imposed assumption that the intergenerational mobility function is linear. Applying our method to the Panel Studies of Income Dynamics (PSID) data, we decisively reject the commonly imposed linearity assumption and find substantial heterogeneity in the IGEs across the population. The IGEs with respect to parental income exhibit a U-shape pattern. Specifically, there is a considerable degree of mobility among the broadly defined middle class, but the children of both high- and low-income parents are more likely to be high- and low-income adults, respectively. Our result suggests that the U.S. is indeed a "land of opportunity", just not for everyone! This result also provides valuable insights into the (intertemporal) Great Gatsby curve, suggesting that a higher level of inequality within one generation may lead to a higher level of social immobility in the next generation in the U.S..

E0467: **A new family of copulas, with application to the estimation of a production frontier system**
Presenter: **Artem Prokhorov**, University of Sydney, Australia
Co-authors: Peter Schmidt, Christine Amsler

A system of equations is considered where one equation is a production function and the other equations are the first order conditions for cost minimization. The equation representing the production function contains a one-sided error that represents technical inefficiency. Also, because the first order conditions for cost minimization will not be satisfied exactly, the corresponding equations contain errors that represent allocative inefficiency. If technical and allocative inefficiency are not independent, we encounter the issue that common copulas do not capture the type of dependence that the economic model implies. What we want is a positive correlation between technical inefficiency and the absolute value of allocative inefficiency, which says that firms that are more technically inefficient have capital/labor ratios that are more in error than more technically efficient firms. The same argument can apply in a non-frontier setting. Even if u and v are standard zero-mean errors (e.g. normal), it may be reasonable to assume that u is correlated with $|v|$ rather than v, reflecting the view that firms that are better at using the correct input ratios also on average produce more output from a given set of inputs.

EO170 Room P4703 RECENT ADVANCES IN FUNCTIONAL AND MULTIVARIATE DATA ANALYSIS Chair: Yuko Araki

E0604: **Functional classification for high dimensional data**
Presenter: **Yuko Araki**, Shizuoka University, Japan

Classification of high dimensional data, such as images or longitudinal data, is considered by using basis expansions with the help of subspace methods. The proposed method is capable of classifying very high dimensional data by using composite basis expansions with sparse principal component analysis. The selection of an appropriate dimensions of decision space are selected by using model selection criteria. The proposed methods are applied to real image data analysis and Monte Carlo simulations are conducted to examine the efficiency of our modelling strategies comparing to other classification methods.

E0514: **Multivariate functional clustering and its application to typhoon data**
Presenter: **Toshihiro Misumi**, Yokohama City University, Japan
Co-authors: Hidetoshi Matsui, Sadanori Konishi

A multivariate nonlinear mixed effects model is proposed for clustering multiple longitudinal data. The advantages of the nonlinear mixed effects model are that it is easy to handle unbalanced data which are highly occurred in the longitudinal study, and that it can account for correlations at a given time point among longitudinal characteristics. The joint modeling for multivariate longitudinal data, however, requires high computational cost because numerous parameters are included in the model. To overcome this issue, we perform a pairwise fitting procedure based on a pseudo-likelihood function. Unknown parameters included in each bivariate model are estimated by the maximum likelihood method along with the EM algorithm, and then the numbers of basis functions included in the model are selected by model selection criteria. After estimating the model, a non-hierarchical clustering algorithm by self-organizing maps is implemented to predicted coefficient vectors of individual specific random effect functions. We present the results of application of the proposed method to the analysis of data of typhoon occurred between 2000-2017 in Asia.

E0362: **Supervised sparse hierarchical components analysis with application to resting-state functional MRI data**
Presenter: **Atsushi Kawaguchi**, Saga University, Japan

Brain functional connectivity is useful for identifying biomarkers that can be used for diagnosis of brain disorders. This connectivity can be measured using resting-state functional Magnetic Resonance Imaging (rs-fMRI). Previous studies were based on a sequential application of the graphical model for network estimation and machine learning for constructing a prediction formula for the outcome (e.g., disease or healthy) from the estimated network. This approach results in a less informative network for diagnosis because the network is estimated independently of the outcome. A regression method with the score from the rs-fMRI based on the supervised sparse hierarchical components analysis (SSHCA) is proposed. SSHCA has a hierarchical structure consisting of the network model (block scores at the individual level) and the scoring model (super scores at the population level). The multiple logistic regression model with super scores as the predictor was used to estimate the diagnostic probability. An advantage of the proposed method is that the outcome-related (supervised) network connection and the multiple scores corresponding to sub-network estimation will be helpful in interpretation. The proposed method was applied to the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and was compared with the existing method based on the sequential approach.

EO318 Room P4704 ADVANCED COMPUTATIONAL METHODS FOR MODELLING COMPLEX SURVIVAL DATA Chair: Yi Niu

E0305: **Bayesian cure-rate survival modeling with spatially structured censoring**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States
Co-authors: Georgiana Onicescu

Spatial geo-referencing of survival models is now established and various different approaches are found. Usually the rate parameter of a survival distribution is defined to have such structure. However the censoring mechanism in survival studies would also display important spatial structure and require sensitive modeling. We examine a Bayesian cure rate model which also includes spatial referencing of the censoring mechanism. An application is made to the analysis of prostate cancer from the SEER registry in Louisiana USA.

E0458: **The Cox proportional hazards cure model in application of disease screening**
Presenter: **Lisha Guo**, South-Central University for Nationalities, China
Co-authors: Xiaoqiong Joan Hu, Yanyan Liu

Motivated by a tuberculosis (TB) study, we consider likelihood-based estimation under the Cox proportional hazards cure model using right-censored times together some covariate entries missing not at random. We demonstrate the existence of the pseudo semiparametric maximum likelihood estimator of the model parameters, and establish the consistency and weak convergence of the estimator. Finite sample performance of the proposed estimation procedure is examined and compared with two conventional approaches. An analysis of the tuberculosis (TB) study data is used to illustrate the application of the statistical approach in disease screening.

E0471: **Some computational methods for cure models**
Presenter: **Yingwei Peng**, Queenś University, Canada

Cure models received a great deal of attention in recent decades and the new development in models and estimation methods in cure models demonstrates their potential applications in cancer research and other fields. To maximize their potential impact in these fields, we focus in this work on computational methods in some cure models, particularly mixture cure models and clustered survival data with a cure fraction. We show some approaches that have potential to make widespread applications of some cure models possible. Illustration with real data are also given.

Wednesday 20.06.2018 08:30 - 09:50 Parallel Session F – EcoSta2018

EO275 Room LT-12 RECENT ADVANCES IN HIGH-DIMENSIONAL NONPARAMETRIC INFERENCE Chair: Miles Lopes

E0252: **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 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.

E0344: **Consistent estimation for partition-wise regression and classification models**
Presenter: **Rex Cheung**, San Francisco State University, United States
Partition-wise models offer a flexible approach for modeling complex and multidimensional data and are capable of producing interpretable results. They are based on partitioning the observed data into regions, each of which is modeled with a simple submodel (similar to the CART modeling). The success of this approach highly depends on the quality of the partition, as too large a region could lead to a non-simple submodel, while too small a region could inflate estimation variance. This project proposes an automatic procedure for choosing the partition (i.e., the number of regions and the boundaries between regions) as well as the submodels for the regions. It is shown that, under the assumption of the existence of a true partition, as well as a set of known significant predictors, the proposed partition estimator is statistically consistent. Features of the proposed methodology are highlighted for both regression and classification problems on synthetic and real data.

E0503: **Empirical likelihood based covariance matrix estimation**
Presenter: **Sanjay Chaudhuri**, National University of Singapore, Singapore
Empirical likelihood based methods are discussed for estimating covariance matrices from data. It is well known that classical estimates cannot be easily modified to include equality based constraints among its entries while keeping it non-negative definite. It is known that empirical likelihood based methods can estimate parameters with additional equality constraints on the expectation of various functionals of the data. We employ empirical likelihood to estimate covariance matrix with equality constraints on its diagonal and off-diagonal entries. We show that the proposed estimates are non-negative definite and does not depend on the underlying distribution of the observations. Such methods can also be extended to rank based constraints and to estimation of high dimensional covariances.

EO322 Room LT-14 STATISTICAL COMPUTING AND OPTIMIZATION Chair: Yoonkyung Lee

E0368: **Computing conditional density of eigenvalues in high-dimension**
Presenter: **Yunjin Choi**, National University of Singapore, Singapore
A method is proposed for evaluating conditional density of eigenvalues of a Wishart matrix in high-dimension. Evaluating the density of eigenvalues involve multi-dimensional integration, while multi-dimensional integration can be computationally challenging especially in high-dimensional setting. This issue has been previously addressed by utilizing approximation of a random matrix kernel and proposed a method for evaluating the marginal distribution of the largest eigenvalue of a Wishart matrix. We extend this approach and propose a method for evaluating the conditional distribution of any k-th eigenvalue with its preceding eigenvalues conditioned. The proposed method can be used for testing the significance of the principal components.

E0389: **Cross validation for penalized quantile regression with a case-weight adjusted solution path**
Presenter: **Yoonkyung Lee**, Ohio State University, United States
Co-authors: Shanshan Tu, Yunzhang Zhu
Cross validation is widely used for selecting tuning parameters in regularization methods, but it is computationally intensive in general. To lessen its computational burden, approximation schemes such as generalized approximate cross validation (GACV) are often employed. However, such approximations may not work well when non-smooth loss functions are involved. As a case in point, approximate cross validation schemes for penalized quantile regression do not work well for extreme quantiles. We propose a new algorithm to compute the leave-one-out cross validation scores exactly for quantile regression with ridge penalty through a case-weight adjusted solution path. Resorting to the homotopy technique in optimization, we introduce a continuous embedding parameter-a case weight for each individual case and decrease the weight gradually from one to zero to link the estimators based on the full data and those with some case deleted. This allows us to design a solution path algorithm to compute all leave-one-out estimators very efficiently from the full-data solution. We show that the case-weight adjusted solution path is piecewise linear in the weight parameter, and using the solution path, we examine case influences comprehensively and observe that different modes of case influences emerge, depending on the specified quantiles, data dimensions and penalty parameter.

E0641: **A surprising connection between single-linkage, graphical lasso, sparse PCA, and other L1 penalized estimators**
Presenter: **Vincent Vu**, The Ohio State University, United States
Statistical sufficiency formalizes the notion of data reduction. In the decision theoretic interpretation, once a model is chosen all inferences should be based on a sufficient statistic. However, suppose we start with a set of methods that share a sufficient statistic rather than a specific model. Is it possible to reduce the data beyond the statistic and yet still be able to compute all of the methods? We will present some progress towards a theory of "computational sufficiency" and show that strong reductions can be made for large classes of penalized M-estimators by exploiting hidden symmetries in the underlying optimization problems. These reductions can (1) enable efficient computation and (2) reveal hidden connections between seemingly disparate methods. As a main example, we will show how the theory provides a surprising answer to the following question: "What do the Graphical Lasso, sparse PCA, single-linkage clustering, and L1 penalized Ising model selection all have in common?"

E0366: **Alternating minimization, proximal minimization and optimization transfer are equivalent**
Presenter: **Jong Soo Lee**, University of Massachusetts Lowell, United States
Proximal minimization algorithms (PMA), majorization minimization (MM), and alternating minimization (AM) are shown to be equivalent. Each type of algorithm leads to a decreasing sequence of objective function. New conditions on PMA are given (the limit of the decreasing sequence of objective function is indeed the infimum of the objective function), which lead to new conditions on AM for the sequence Phi to converge to its infimum. These conditions can then be translated into the language of MM. Examples are given of each type of algorithm and some open questions are posed.

EO075 Room P4703 MIXTURE MODELS FOR CENSORED AND LONGITUDINAL DATA Chair: Victor Hugo Lachos Davila

E0275: **Clustering multi-outcome longitudinal data via finite mixtures of multivariate t linear mixed models**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan

The issues of model-based clustering and classification of longitudinal data have received increasing attention in recent years. A finite mixture of multivariate t linear mixed-effects model (FM-MtLMM) is presented for analyzing longitudinally measured multi-outcome data arisen from more than one heterogeneous sub-population. The motivation comes from a cohort study of patients with primary biliary cirrhosis (PBC), where the interest is in classifying new patients into two or more prognostic groups on the basis of their longitudinally observed bilirubin and albumin levels. The proposed FM-MtLMM offers robustness and flexibility to accommodate fat tails or atypical observations contained in one or several of the groups. An efficient alternating expectation conditional maximization (AECM) algorithm is employed for computing maximum likelihood estimates of parameters. Practical techniques for clustering of multivariate longitudinal data, estimation of random effects, and classification of future patients are also provided. The methodology is illustrated by analyzing Mayo Clinic PBC sequential data and a simulation study.

E0579: **Analysis of longitudinal interval censored data using finite mixture of multivariate Student-t distributions**
Presenter: **Christian Galarza**, State University of Campinas, Brazil
Co-authors: Victor Hugo Lachos Davila

Mixture models are based on the assumption of normality (symmetry) and thus are sensitive to outliers, heavy-tailed and skewness. Besides, these kind of data can be subject to some upper and/or lower detection limits because of the restriction of experimental apparatus. For such data structures, we present a proposal to deal with these issues simultaneously by propose an interval censored regression based on finite mixtures of multivariate Student-t distributions. This approach allows us to model data with great flexibility, accommodating multimodality, heavy tails and 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 presented. Results obtained from the analysis of a part of Signal Tandmobiel data, which contains observed intervals of teeth emergence for 4430 Flemish children resulting from a longitudinal project, is reported to demonstrate the effectiveness of the proposed methodology.

E0361: **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 multi-modality, 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.

EC291 Room G4302 CONTRIBUTIONS IN TIME SERIES Chair: Qiying Wang

E0357: **Anomaly detection in clustered multiple time series**
Presenter: **Ruby Jean Ocenar**, University of the Philippines Diliman, Philippines
Co-authors: Erniel Barrios

A nonparametric test procedure based on the bootstrap for detecting structural change in clustered multiple time series data is proposed. Simulation studies indicate that the test is correctly sized in stationary time series and that when locally, the time series does not manifest non-stationarity. The test is powerful specially if all time series exhibit structural change, and when the change happened towards the more recent realization of the time series. Power is still high even when structural change occurred only in some clusters provided that change occurred in greater magnitude.

E0739: **A test for serial dependence using neural networks**
Presenter: **Jinu Lee**, King's College London, United Kingdom
Co-authors: George Kapetanios

Testing serial dependence is central to much of time series econometrics. A number of tests that have been developed and used to explore the dependence properties of various processes. We build on recent work on nonparametric tests of independence. We consider a fact that characterises serially dependent processes using a generalisation of the autocorrelation function. Using this fact we build dependence tests that make use of neural network based approximations. We derive the theoretical properties of our tests and show that they have superior power properties. Our Monte Carlo evaluation supports the theoretical findings. An application to a large dataset of stock returns illustrates the usefulness of the proposed tests.

E0799: **Stationary and nonstationary time series models for categorical data**
Presenter: **Lionel Truquet**, ENSAI, France
Co-authors: Konstantinos Fokianos

Finite-state Markov chains are of limited use in the modeling of dependent data because the number of parameters grows exponentially with the order of the chain. To get more parsimonious models, some logistic type autoregressions that involve a latent process are considered in the literature. We will present optimal conditions for stationarity and ergodicity of such processes. In the nonstationary case, we will also discuss locally stationary versions of these models as well as some asymptotic results for statistical inference. Our results are highly based on coupling techniques for a general class of finite-state processes, the chains with complete connections.

E0670: **Response surface models for the Elliott-Rothenberg-Stock and Leybourne unit root tests**
Presenter: **Christopher Baum**, Boston College, United States
Co-authors: Jesus Otero

Response surface coefficients are presented for a large range of quantiles of the Elliott, Rothenberg and Stock DF-GLS and Leybourne ADFmax unit root tests, for different combinations of number of observations T, and lag order in the test regressions, p. The lag order can be either specified by the user or endogenously determined from SIC or AIC criteria or chosen by the Ng-Perron general-to-specific procedure. The critical values depend on the method used to select the number of lags, and vary considerably over p and the method used. The user-contributed Stata commands ersur and adfmaxur are presented and their use illustrated with empirical examples that consider how the inference varies depending on the method

used to determine lag length.

EC292 Room G4701 CONTRIBUTIONS IN MULTIVARIATE METHODS Chair: Donatello Telesca

E0783: **Structural equation modeling and canonical correlation analysis**
Presenter: **Zhenqiu Laura Lu**, University of Georgia, United States
Co-authors: Fei Gu
Canonical Correlation Analysis (CCA) is a generalization of multiple correlation that examines the relationship between two sets of variables. Traditional methods apply spectral decomposition to obtain canonical correlations and canonical weights. It has also been previously provided the asymptotic distribution of the canonical weights under normality assumption. We propose a new approach by using Structural Equation Modeling (SEM) approach to canonical correlation analysis. Mathematical forms are presented to show the equivalence among these models. Popular SEM software such as Lavaan, Mplus, EQS are demonstrated to illustrate the application. The weight matrix is obtained as the inverse of the loading matrix. And the variance or standard errors of weights are calculated through the delta method. The results obtained from SEM approach are compared with those obtained from traditional CCA approach and also from the Andersons (2003) formula. Advantages of this new approach include that (1) it provides both canonical correlations and the covariances of canonical variates, (2) it is very practically and flexible because existing SEM software can be used. Related issues are also discussed in the last section.

E0729: **Classification with imperfect training labels**
Presenter: **Timothy Cannings**, University of Southern California, United States
Co-authors: Yingying Fan, Richard Samworth
The effect of imperfect training data labels on the performance of classification methods is studied. In a general setting, where the label errors occur at random and the probability that an observation is mislabelled depends on the feature vector and true label of the observation, we bound the average misclassification error of an arbitrary classifier trained with imperfect labels. Furthermore, under conditions on the labelling error probabilities, we derive the asymptotic properties of the popular k-nearest neighbour ($k$nn), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. The $k$nn and SVM classifiers are robust to imperfect training labels, in the sense that these methods are Bayes consistent. In fact, we see that in some cases imperfect labels can improve the performance of these methods. On the other hand, the LDA classifier is not robust to label noise, unless the prior probabilities of the classes are equal. Finally, our theoretical results are demonstrated via a simulation study.

E0620: **VAR estimation impacts on frequency causality measures**
Presenter: **Thibault Soler**, University Paris 1 – Pantheon-Sorbonne, France
Co-authors: Philippe De Peretti, Christophe Chorro, Emmanuelle Jay
Granger non-causality tests have received a great deal of attention over recent years. In time domain, tests have been extended to deal with co-integration, near unit roots, mixtures of variables with different integration and structural breaks. In the frequency domain, methodologies have been developed to quantify causal relationships. These approaches are two-step ones consisting in first estimating a Vector AutoRegressive model (VAR), and then computing coherence of transfer function. Nevertheless, the first step may be incorrectly performed for many reasons: incorrect model order selection, insufficient number of observations regarding the number of series, the omission of zero-lag effect, etc., then we are likely to have cascading errors and flawed results. The goal is, therefore, twofold: first, evaluate the impact of a non-efficient VAR estimation on causality measures in the frequency domain, then suggest an efficient two-step methodology to overcome cascading errors in estimating a multivariate model. The first step consists in using the modified backward in time selection method (mBTS), and the second one makes use of the top-down strategy (TD). The impact of error diagonal covariance, time series lengths, and instantaneous effects on causality measures is studied first on simulated data, and then on financial data. The results are compared between our methodology (mBTS-TD) and existing methods of variable subset selection (mBTS, TD, bottom-up (BU)).

E0748: **Selecting the number of maximum autocorrelation factors**
Presenter: **Souveek Halder**, Australian National University, Australia
Dimension reduction is a popular technique in statistics, used to transform high dimensional data into a low dimensional space. Principal Component Analysis (PCA) is one of the classical methods that is popularly used by researchers for dimension reduction. However, a major drawback of PCA is that it only provides the best linear approximation for a high dimensional dataset. Maximum Autocorrelation Factors (MAF) was developed as an alternative to PCA by Switzer and Green in the 1980s. One significant challenge in using MAF for dimension reduction is to determine how many factors to retain or, in other words, how much dimension reduction is to be done. In most cases the choice is made using some ad-hoc method such as an autocorrelation scree-plot or a bootstrap based technique. However, these methods cannot be considered best in all situations. Thus, our focus is on tests that deal with eigenvalues, as computing MAF is equivalent to solving an eigenvalue/eigenfunction problem. We evaluate the performance of Conditional Singular Value test (CSV), a signal-to-noise ratio test, the Pseudorank method, and a likelihood ratio test.

EC295 Room LT-11 CONTRIBUTIONS IN FORECASTING Chair: Henghsiu Tsai

E0778: **Forecasting multivariate realized volatility using time varying coefficient models**
Presenter: **Laleh Tafakori**, University of Melbourne, Australia
Co-authors: Hans Manner, Bing Liu
Modelling- and forecasting- multivariate realised volatility plays an indispensable role for option pricing, portfolio allocation and risk management. The existing models for realised volatility may perform well in-sample but in general their out-of-sample forecasts are often biased. We aim to build a model for realised volatility with improved forecasting performance by accounting for the fact that the multivariate realised covariances are the only estimates of the true variance and by introducing time varying parameters. With its more accurate forecasting, our model holds the promise to empirically more accurate pricing models and improved financial decision-making.

E0766: **Identification robust predictive ability testing**
Presenter: **Andrea Naghi**, Erasmus University, Netherlands
The aim is to analyze the predictive accuracy evaluation of models that are strongly identified in some part of the parameter space but non-identified or weakly identified in another part of the parameter space. We show that when comparing the predictive ability of models that might be affected by identification deficiencies, when the parameter estimation error is negligible, the null distribution of out-of-sample predictive ability tests is not well approximated by the standard normal distribution. As a result, employing a standard (strong) identification critical value can lead to misleading inference. We propose methods to make the out-of-sample predictive ability tests robust to identification loss. These methods use a different critical value than the standard one and include: a least-favorable critical value and a data dependent critical value. In settings where the parameter error is non-negligible, it is shown that the asymptotic distribution of the usual predictive ability test is standard, even when one allows for the model(s)to be only semi-strongly identified.

E0790: **Forecasting macroeconomic series by unobserved component models with ARMA-SV errors**
Presenter: **Bo Zhang**, Australian National University, Research school of Economics, Australia

An autoregressive moving average component with stochastic volatility is introduced into the unobserved component model. A transformation in a stacked matrix form of the model is conducted for the posterior fast simulation, and a recently developed precision-based algorithm, particularly for the unobserved component model, is adopted for analyzing the serially dependent errors. The proposed model is then used to study macroeconomic time series in the United States. It is found that the proposed new model provides good full sample simulation for a large part of the macroeconomic variables, and it can improve both point forecast and interval forecast performance of these variables across different horizons.

E0716: **Probabilistic forecasts in hierarchical time series**
Presenter: **Puwasala Gamakumara**, Monash University, Australia
Co-authors: Anastasios Panagiotelis, George Athanasopoulos, Rob Hyndman

Forecasting hierarchical time series has been of great interest in many applications. While there is a rich literature on hierarchical point forecasting, we focus on a probabilistic hierarchical framework. We initially provide a theoretical foundation for probabilistic forecast reconciliation by considering the aggregation structure of a hierarchy. We observe that the trace minimization (MinT) approach in producing optimal point forecasts, is also generating optimal probabilistic forecasts under Gaussianity. We further relax the Gaussian assumption and propose a novel, non-parametric approach. This involves first simulating future sample paths of the whole hierarchy using bootstrapped training errors and then reconciling these sample paths so that they become coherent. We evaluate both the MinT Gaussian and non-parametric bootstrap approaches via extensive Monte Carlo simulations.

EC296 Room LT-18 CONTRIBUTIONS IN COMPUTATIONAL AND NUMERICAL METHODS Chair: Berwin Turlach

E0732: **Hamiltonian Monte Carlo using efficient importance sampling: Exploring the speed-accuracy trade-off**
Presenter: **Kjartan Kloster Osmundsen**, University of Stavanger, Norway
Co-authors: Tore Selland Kleppe, Roman Liesenfeld

The joint posteriors of latent variables and parameters in Bayesian hierarchical models often have strong nonlinear dependencies, and thus making them challenging targets for standard Markov chain Monte Carlo methods. Pseudo- marginal methods are able to effectively explore such target distributions, by integrating out the latent variables and directly targeting the marginal posteriors of the parameters. The combination of Efficient Importance Sampling (EIS) for integrating out latent variables and recently proposed pseudo-marginal Hamiltonian Monte Carlo for sampling from parameter marginal is explored. The methodology is shown to be highly efficient in the context of state space models.

E0737: **Speed-up of bootstrap computation of the covariance matrix of MLEs from incomplete data**
Presenter: **Masahiro Kuroda**, Okayama University of Science, Japan
Co-authors: Yuichi Mori

The bootstrap is a most useful method to compute the covariance matrix of maximum likelihood estimates (MLEs) of parameters given a statistical model. In the bootstrap computation, we generate a bootstrap sample by randomly sampling with replacement from observed data and compute the MLEs using this sample. After repeating the procedure B times, we can obtain the covariance matrix of the MLEs from the B MLEs. When applying the bootstrap to incomplete data, we require to add an iterative computation step of finding MLEs to the bootstrap procedure. The EM algorithm is used in the MLE computation step and is applied to each of B bootstrap samples. Then the bootstrap computation for incomplete data takes long computation time due to the slow convergence of the EM algorithm. In order to reduce the bootstrap computation cost, we provide a simple acceleration algorithm for speeding up the convergence of the EM algorithm. Numerical experiments examine the performance of the speed-up of the bootstrap computation using the accelerated EM algorithm for incomplete data.

E0780: **Numerical computation of the higher order central moments of the multivariate normal distribution**
Presenter: **Fumiyasu Komaki**, The University of Tokyo, Japan

The higher order central moments of the multivariate normal distribution naturally appear in a problem of Bayesian prediction. Although a general formula (Isserlis' theorem) for the central moments of the multivariate normal distribution is widely known, it is not suitable for numerical evaluation of them. We investigate a method to evaluate the higher order central moments of the multivariate normal distribution by using MCMCMC. A class of discrete distributions closely related to the higher order moments is introduced. Applications of the method to Bayesian prediction are discussed.

E0725: **Saddlepoint adjusted inversion of characteristic functions for likelihood optimisation**
Presenter: **Berent Aanund Stroemnes Lunde**, University of Stavanger, Norway
Co-authors: Tore Selland Kleppe, Hans Skaug

For certain types of statistical models, the characteristic function is available in closed form, whereas the probability density function will have an intractable form, typically as an infinite sum of probability weighted densities. Important examples include solutions of stochastic differential equations with jumps, the Tweedie model, and certain mixture models. Likelihood optimisation, using inversion of the characteristic function, is made difficult by possible multi-modality of the density, which it is shown renders the unimodal saddlepoint approximation (SPA) useless. Direct numerical integration techniques will only work up-until a constant, which creates numerical problems of taking logarithms of the approximation to the density in the tails, something very intractable for optimisation routines. As a solution, the integrand of the problem is optimized for "well behaviour" under numerical inversion, much like the original SPA, creating a SPA weighted numerical inversion technique that is exact over the whole domain. The routine is computationally stable under optimisation, while also being much faster than ordinary SPA renormalisation routines, along with the tractable property of being exact. The method is applied to likelihood estimation of jump diffusion models, the Tweedie model, and mixture models and is also empirically seen to be stable, efficient, and accurate.

EC303 Room P4701 CONTRIBUTIONS IN APPLIED STATISTICS AND ECONOMETRICS Chair: Feng Chen

E0677: **Optimal model averaging estimation for correlation structure in generalized estimating equations**
Presenter: **Jingli Wang**, National university of Singapore, Singapore

Longitudinal data analysis requires a proper estimation of the within-cluster correlation structure in order to achieve efficient estimates of the regression parameters. When applying likelihood-based methods one may select an optimal correlation structure by the AIC or BIC. However, such information criteria are not applicable for estimating equation based approaches. We develop a model averaging approach to estimate the correlation matrix by a weighted sum of a group of patterned correlation matrices under the GEE framework. The optimal weight is determined by minimizing the difference between the weighted sum and a consistent yet inefficient estimator of the correlation structure. The computation of our proposed approach only involves a standard quadratic programming on top of the standard GEE procedure and can be easily implemented in practice. We provide theoretical justifications and extensive numerical simulations to support the application of the proposed estimator. A couple of well-known longitudinal data sets are revisited where we implement and illustrate our methodology.

E0761: **Application of multi-domain clustering to C. elegans neural network analysis**
Presenter: **Stephen Wu**, The Institute of Statistical Mathematics, Japan
Co-authors: Ye Liu, Michael Ng, Mirai Tanaka
Whole-brain imaging of C. elegans allows neuroscientist to access the full neural network activity of a single worm under different stimulations, which will be an important step for understanding brain activities. However, the noisy nature of the images makes it difficult to extract meaningful activity patterns using conventional clustering methods. A typical statistical solution is to increase the number of worm samples in order to suppress the influence from the noise. In order to take the full advantage of multiple data sets, we formulate the neural network analysis of multiple worms as a multi-domain clustering problem. We construct an undirected graph for each worm to represent the correlation of the neural activities between neurons. The preliminary results of our multi-domain clustering method show interesting biological meanings that may guide the future experiment of the C. elegans research.

E0760: **Time-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 reveal broader 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, yet, their amplitudes vary significantly in time.

E0168: **Identifying leverage effect in intra-day volatility pattern: Toward a functional data analysis**
Presenter: **Ping Chen Tsai**, Southern Taiwan University of Science and Technology, Taiwan
Estimating intra-day volatility pattern (IVP) consists of obtaining the weights of volatility over a high dimension of intra-day intervals. The natural ordering of intra-day intervals, however, renders an interpretation of IVP as functional data. A new stylized fact of volatility is documented by identifying leverage effect in the U-shape intra-day volatility pattern, with days following negative and positive returns presenting different such patterns. Functional forms for the IVPs are then determined by non-parametric methods including interpolation and smoothing. The obtained functional forms of IVP can be incorporated into the estimation of realized variance and realized bi-power variation, which in turn has an important impact on testing for jumps in prices. A Monte Carlo study shows that the overall size and power of jump tests are improved after accounting for the leverage effect in IVP.

EC298 Room P4704 CONTRIBUTIONS IN STATISTICAL MODELLING Chair: Ray-Bing Chen

E0754: **Joint models of longitudinal and survival data with heteroscedastic random effects covariance matrix**
Presenter: **Jaewoong Joo**, Sungkyunkwan University, Korea, South
Co-authors: Donguk Kim, Keunbaik Lee
In clinical trials, longitudinal data are collected repeated over time. However, these data have a tendency to show many missing values because of censoring problems, and these missing data are not totally negligible. In order to cope with this, a joint model of longitudinal data with repetitive measurements and competing risks survival data is considered. The random effects in the model are used to account for association between the longitudinal and survival submodels. However, estimation of the random effects covariance matrix is not easy because the matrix is high-dimensional and the estimated covariance matrix should be positive definite. We consider modified Cholesky decomposition (MCD) to overcome these limitation of the random effects covariance matrix in the joint model. In addition, the estimated covariance matrix using MCD can be heteroscedastic. We then used our proposed model to analyze CD4 cell count data.

E0764: **Full hairpin copulas are not factorizable**
Presenter: **Songkiat Sumetkijakan**, Chulalongkorn University, Thailand
Co-authors: Tanes Printechapat
Recently, a subclass of singular bivariate copulas, called non-atomic copulas, were introduced. We show that hairpin copulas are non-atomic. Furthermore, we prove that full hairpin copulas are not factorizable, that is they cannot be written as $L*R$ where L and R are left and right invertible copulas.

E0743: **Estimation of factor scores: Comparing parametric and non-parametric approaches.**
Presenter: **Tim Fabian Schaffland**, University of Tuebingen, Germany
Co-authors: Stefano Noventa, Augustin Kelava
Estimation of factor scores in latent variable models has repeatedly attracted the interest of researchers for decades. Already in 1935 Thurstone proposed the regression method, and in 1937 Bartlett suggested his well-known approach. Still today, factor score estimation and their properties, for example the bias of their moments, raise debate and interest. We will compare the Bartlett estimator, the regression method, the least square estimation, and one new approach which makes no distributional assumptions on the latent variables. Factor scores are estimated by combining the empirical CDF and the independence assumption between the measurement errors and the latent factors. This results in factor score estimates that in theory could consistently replicate the true joint distribution of the latent variables and the measurement error. In a simulation study we vary the (multivariate) distribution of the underlying factors and examine the performance of the different approaches in recovering the first four moments of the joint distribution of the latent variables. Additionally, the influence of the factor loadings on the estimation is investigated. Two different ways of estimating the factor loadings are used as well as the true values of the loadings. We conclude with the implications and recommendations for factor score estimation in an applied context.

E0727: **Modified check loss for efficient model selection in quantile regression**
Presenter: **Yoonsuh Jung**, Korea University, Korea, South
Co-authors: Steven MacEachern, Hang Kim
Check loss function is used to define quantile regression. In the prospect of cross-validation, it is also employed as a validation function when true distribution is unknown. However, our empirical study indicates that the validation with check loss often leads to choose an over estimated fits. We suggest a modified or L2-adjusted check loss which rounds the sharp corner in the middle of check loss. It has a large effect of guarding against over fitted model in some extent. Through various simulation settings of linear and non-linear regressions, the improvement of check loss by quadratic adjustment is empirically examined. This adjustment is devised to shrink to zero as sample size grows.

EG025 Room B4302 CONTRIBUTIONS ON REGRESSION AND APPLICATIONS Chair: Hung-pin Lai

E0261: **Volatility forecasting using the HAR and lasso-based models: An empirical investigation**
Presenter: **Xingzhi Yao**, Lancaster University, United Kingdom
Co-authors: Marwan Izzeldin

The aim is to compare the performance of various least absolute shrinkage and selection operator (Lasso) based models in forecasting future log realized variance (RV) constructed from high-frequency returns. We conduct a comprehensive empirical study using the SPY and 10 individual stocks selected from 10 different sectors. In an in-sample analysis, we provide evidence for the invalidity of the lag structure implied by the heterogeneous autoregressive (HAR) model in volatility forecast. In our out-of-sample study, the best forecasting performance is usually provided by the Lasso-based model and the idea of forecast combination tends to improve the forecasting accuracy of the Lasso-based model. Among all models of interest, the ordered Lasso AR using the forecast combination serves as the top performer most frequently in forecasting RV and its improvements over the HAR model are, in most cases, significant over monthly horizons. In line with the existing study, the superiority of the Lasso-based models is more evident in a larger forecasting window size.

E0714: **Semiparametric efficient estimators in heteroscedastic error models**
Presenter: **Mijeong Kim**, Ewha Womans University, Korea, South
Co-authors: Yanyuan Ma

In the mean regression context, several frequently encountered heteroscedastic error models are considered where the regression mean and variance functions are specified up to certain parameters. An important point we note through a series of analyses is that different assumptions on standardized regression errors yield quite different efficiency bounds for the corresponding estimators. Consequently, all aspects of the assumptions need to be specifically taken into account in constructing their corresponding efficient estimators. The relation between the regression error assumptions and their respectively efficiency bounds is clarified under the general regression framework with heteroscedastic errors. Our simulation results support our findings; we carry out a real data analysis using the proposed methods where the CobbDouglas cost model is the regression mean.

E0801: **Forecast evaluations under asymmetric loss functions**
Presenter: **Pin Ng**, Northern Ariziona University, United States
Co-authors: Zhijie Xiao, Kemal Guler

Forecasts are pervasive in all areas of applications in business and daily life. Hence evaluating the accuracy of a forecast is important for both the generators and consumers of forecasts. On measuring the accuracy of a past forecast, the aim is to illustrate that the summary statistics used should match the loss function that was used to generate the forecast. If there is strong evidence that an asymmetric loss function has been used in the generation of a forecast, then a summary statistic that corresponds to that asymmetric loss function should be used in assessing the accuracy of the forecast instead of the popular root mean square error or mean absolute error. On testing the optimality of the forecasts, it is demonstrated how the quantile regressions set in the prediction-realization framework of Mincer and Zarnowitz can be used to recover the unknown parameter that controls the potentially asymmetric loss function used in generating the past forecasts. Finally, the prediction-realization framework is applied to the Federal Reserves economic growth forecast and forecast sharing in a PC manufacturing supply chain. It is found that the Federal Reserve values overprediction approximately 1.5 times more costly than underprediction. It is also found that the PC manufacturer weighs positive forecast errors (under forecasts) about four times as costly as negative forecast errors (over forecasts).

E0808: **Active predictor detection by controlling the false discovery rate**
Presenter: **Yuanyuan Lin**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Wenlu Tang, Jinhan Xie

In modern scientific discoveries, important variables identification in analyzing high dimensional data is intrinsically challenging, especially when there are complex relationships among predictors. Without any specification of a regression model, we introduce an association statistic based on quantiles to identify influential predictors, which is flexible to capture a wide range of dependence. The asymptotic null distribution of the proposed statistic is established under mild conditions. Moreover, a multiple testing procedure is advocated to simultaneously test the independence between each predictor and the response variable in ultra-high dimensionality. It is computationally efficient as no optimization or resampling is involved. We prove its theoretical properties rigorously and justify the proposal asymptotically controls the false discovery rate. Numerical studies including simulation studies and real data analysis contain supporting evidence that the proposal performs reasonably well in practical settings.

EG329 Room LT-13 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS Chair: Yang Shen

E0712: **Testing for serial correlation of unknown form using signed path dependence**
Presenter: **Fabio Dias**, University College London, United Kingdom
Co-authors: Franz Kiraly, Gareth Peters

Whilst several tests for serial correlation in financial markets have been proposed and applied successfully in the literature, such tests provide rather limited information to construct predictive econometric models. This gap is filled by providing the following contributions: (i) a formal definition of signed path dependence based on how the sign of cumulative innovations for a given lookback window correlates with the future cumulative innovations at a given forecast horizon; (ii) theoretical results validating the definition on well-known model classes; (iii) a formal inference procedure to detect serial correlation of unknown form based on a hypothesis testing formulation of signed path dependence; (iv) experiments on synthetic data validating the test formulation via type I and type II error curves as functions of the significance threshold; (v) an application of the test on observed returns of global equity indices and currencies; (vi) a predictive econometric model for future market returns based on the concept of signed path dependence; and (vii) an application of this model as a profit-seeking trading strategy. It is found strong evidence of serial correlation of unknown form on equity markets, being statistically significant and economically significant even when trading costs are present.

E0722: **Equity risk factors for the long and short run: Pricing and performance at different frequencies**
Presenter: **Patrick Tuijp**, Ortec Finance / University of Amsterdam, Netherlands
Co-authors: Terri van der Zwan, Erik Hennink

The empirical relations between macroeconomic factor models and Fama-French-type factor models are investigated, and their performance in explaining stock returns are compared at different frequencies through spectral analysis. The analysis considers both long-term variation and business cycle variation. The results show that Fama-French-type factors are more suitable than macroeconomic factors to explain the cross-section of equity returns in terms of R2 at multiple frequencies. The risk premia vary in magnitude over time and horizon. The value factor seems to be the most important factor in explaining the cross-section of equity returns on all frequencies. Regarding the macroeconomic factors, the industrial production factor, the credit spread and the term spread seem to play a bigger role on lower frequencies. Interestingly, we see that at lower frequencies macroeconomic factors and Fama-French factors seem to contain related information, while this is not the case at higher frequencies.

E0759: **A dynamic quantile model for bond pricing**
Presenter: **Frantisek Cech**, UTIA AV CR vvi, Czech Republic
Co-authors: Jozef Barunik
A dynamic quantile model for bond pricing is introduced with an agent who values securities by maximizing quantile level of her utility function. The transition from traditional to quantile preferences allows us to study the pricing of the term structure of interest rates by the economic agents differing at their risk aversion. Moreover, the framework is robust to fat tails commonly observed in the empirical data. In the application, we focus on the quantile pricing of the two, five, ten and thirty years US and German government bonds. For the analysis, we use flexible quantile regression framework which is applied over highly liquid bond futures contract from the Chicago Board of Trade and EUREX exchanges.

E0782: **Portfolio strategies with optimal investment to derivatives**
Presenter: **Tomas Tichy**, VSB-TU Ostrava, Czech Republic
Portfolio selection strategies with options are proposed. We generally assume that the returns follow Markov processes that are approximated with proper Markov Chains. Then, we preliminary examine all the American options with underlying the components of the Dow Jones Industrial index and we discuss different portfolio strategies based either on hedging the risk of the underlying, or on optimizing proper expected first passage times of the wealth at some benchmark levels. We focus on an empirical comparison among different hedging and timing portfolio selection strategies. Specifically, we compare the ex post wealth obtained by optimizing new performance strategies based either on hedging of the portfolio risk properly or optimizing the average first passage time of the wealth at some benchmark levels.

EG033 Room LT-17 CONTRIBUTIONS ON STRUCTURAL BREAKS AND CHANGE ANALYSIS Chair: Bonsoo Koo

E0302: **Abrupt change in mean using block bootstrap and avoiding variance estimation**
Presenter: **Michal Pesta**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Barbora Pestova
Sequences of weakly dependent observations that are naturally ordered in time are considered. Their constant mean is possibly subject to change at most once at some unknown time point. The aim is to test whether such an unknown change has occurred or not. The change point methods presented here rely on ratio type test statistics based on maxima of the cumulative sums. These detection procedures for the abrupt change in mean are also robustified by considering a general score function. The main advantage of the proposed approach is that the variance of the observations neither has to be known nor estimated. The asymptotic distribution of the test statistic under the no change null hypothesis is derived. Moreover, we prove the consistency of the test under the alternatives. A block bootstrap method is developed in order to obtain better approximations for the test's critical values. The validity of the bootstrap algorithm is shown. The results are illustrated through a simulation study, which demonstrates computational efficiency of the procedures. A practical application to real data is presented as well.

E0689: **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 derive the asymptotics of a generalized least squares (GLS) estimator of the structural break date in the time series models with a single break in level and/or trend and stationary errors. The asymptotic distribution theory can be readily applied for testing and inference. It is found that the GLS, ordinary least squares (OLS) and GLS quasi-differencing (GLS-QD) break date estimators are asymptotically equivalent. The common asymptotic distribution of these three estimators captures the asymmetry and bimodality often observed in finite samples, and delivers good approximations in general settings. As the GLS estimator relies on the unknown inverse autocovariance matrix, we construct feasible GLS (FGLS) estimators using a consistent estimator of the inverse matrices. Monte Carlo studies show finite sample gains of the FGLS estimators when there is a strong serial correlation. Furthermore, we propose three novel constructions of confidence sets by using the FGLS break date estimators. The confidence sets are based on either a pivotal quantity or the inversion of multiple likelihood-ratio tests. The asymptotic critical value does not depend on nuisance parameters. We find that our proposed methods have fairly accurate coverages and short lengths in various simulations. When there are persistent errors and small break sizes, one of our suggested confidence sets yields good coverage and relatively short length consistently.

E0733: **Structural change and the problem of phantom break locations**
Presenter: **Yao Rao**, The University of Liverpool, United Kingdom
It is well known, in structural break problems, that it is much easier to detect the existence of a break in a data set than to determine the location of such a break in the sample span. The aim is to investigate why, in the context of Gaussian linear regressions, using a decision theory framework. The nub of the problem, even for moderately sized breaks, is that the posterior probability distribution of the possible break points is usually not very informative about the true break location. Hence, even a locally optimal break location procedure, as introduced here, is ineffective. In the regression context, it turns out to be quite common, indeed the norm, for break location procedures to misidentify the true break position up to 100% of the time. Unfortunately too, the magnitude of the difference between the miss-identified and true break locations is usually not small.

E0336: **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.

EG056 Room P4302 CONTRIBUTIONS IN MODELLING FINANCE DATA AND RISK ASSESSMENT Chair: Kaijian He

E0705: **Stochastic modelling of ambient air quality and pricing of air pollution derivatives**
Presenter: **Anders Sleire**, University of Bergen, Norway
Poor air quality in densely populated urban areas is a significant health concern, affecting millions of people globally. In recent years, there has also been increased focus on the impact of air pollution on business and industry. Episodes with extreme pollution levels occur in large metropolitan areas, changing the day-to-day activities of the population. Businesses may suffer financial losses through altered consumer behaviour, or directly from government imposed restrictions aiming to reduce emissions. We introduce air pollution derivatives as instruments for hedging against financial pollution risk. Building upon weather derivatives theory, we design contracts whose payoff depend on publicly available air quality data. The degree of pollution is typically assessed by measuring concentration of key pollutants, such as particulate matter, ground-level ozone, carbon monoxide, sulfur dioxide, lead, and nitrogen dioxide. Results can be communicated to the public on a standardized scale, such as the widely-adopted Air Quality Index (AQI). We develop stochastic models able to capture the seasonality, time-varying volatility, and jumps present

in reported particulate matter AQI for a group of major Asian cities. The models are used to price options with AQI-based indexes as settlement references. Some practical use cases are also presented and discussed.

E0700: **Evaluation and analysis of the value of German real estate following the financial crisis of 2007**
Presenter: **Chong Dae Kim**, Technical University of Cologne, 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.

E0693: **Dynamic cross-sectional copula factor model**
Presenter: **Ziyi Wang**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Mike So
Correlation analysis has been an important component of financial risk analysis. However, the nonlinear dependence among financial returns and time-varying features haven not been fully captured by existing models. By incorporating market factors under the CAPM model, we propose a new cross-sectional vine copula factor model to better capture the dependence among financial returns. Vine decomposition is applied to estimate conditional dependence by expressing a high-dimensional distribution by linking the financial returns to the market factors and linking the market factors by copula functions. With the modeling of the marginal distribution of returns using a GARCH-t structure, the proposed model can capture non-linear and non-monotonic dependence while accounting for heteroscedasticity in financial returns. The computation burden due to high-dimensionality now concerns only the number of market factors, regardless of the dimension of financial returns. Simulation study is performed to illustrate that our methodology works in high-dimensional situations. An empirical study with multiple financial time series is also conducted to illustrate this new model.

E0481: **Dependence structure between Chinese Shanghai and Shenzhen stock market based on copulas and cluster analysis**
Presenter: **Hao Wang**, Jilin University, China
A copula can fully characterize the dependence of multiple variables through the structure and the corresponding dependence measure. The aim is to present a general approach to study the dependence structure of a high-dimensional financial market based on copulas and hierarchical cluster. The idea is to use vine-copula and pair-copula constructions to show the whole dependence structure of a high-dimensional stock market portfolio and to use hierarchical clustering algorithm to group the assets listed in the stock market for portfolio construction. In practice, the method will be used to check whether there is dependence between the dependence of the two Chinese stock markets, namely Shanghai and Shenzhen. Firstly, the copula and the hierarchical cluster via tail dependence and non linear- correlation measure would be fitted for each market. Then, suitable comparisons will be performed.

Wednesday 20.06.2018 10:20 - 12:25 Parallel Session G – EcoSta2018

EI008 Room LT-18 BAYESIAN MODELING FOR COMPLEX STRUCTURES Chair: Igor Pruenster

E0151: **Bayesian method for causal inference in spatially-correlated multivariate time series**
Presenter: **Subhashis Ghosal**, North Carolina State University, United States
Measuring the causal impact of an advertising campaign on sales is important for advertising companies. We propose a novel Bayesian method to infer causality which can also detect weak impacts. We compare two posterior distributions of a latent variable—one obtained by using the observed data from the test stores and the other one obtained by using the data from their counterfactual potential outcomes. The counterfactual potential outcomes are obtained from the data of synthetic controls given by a sparse linear combination of sale figures at many control stores over the causal period. We use a multivariate structural time series model to capture the spatial correlation between test stores. Stationarity is imposed on the local linear trend of the model to prevent the prediction intervals from being explosive. A two-stage algorithm is proposed to estimate the parameters of the model. In Stage 1, a modified EMVS algorithm is applied to select control stores. In Stage 2, an MCMC algorithm is used to obtain the samples of the rest parameters. We present extensive simulation results to show the effectiveness of the proposed method. The new method is applied to measure the causal effect of an advertising campaign for a consumer product sold at stores of a large national retail chain.

E0554: **Bayesian nonparametric inference with heterogeneous data**
Presenter: **Antonio Lijoi**, Bocconi University, Italy
In the last few years, the analysis of non-exchangeable data has attracted considerable interest in the Bayesian nonparametric literature. The aim is to discuss a model based on dependent discrete random probability measures that may be used to deal with data that are generated under different, though related, experimental conditions. Predictive distributions and posterior characterizations will be presented, along with algorithms that allow to determine approximate Bayesian inferences of interest. The discussion will be completed by illustrations with simulated and real data.

E0456: **A dynamic admixture Poisson process analysis of neuronal spike trains**
Presenter: **Surya Tokdar**, Duke University, United States
The brain is able to encode multiple simultaneous stimuli and segment them into objects, but the neural computing behind this complex operation of great relevance to computational and cognitive neuroscience remains poorly understood. Presently lacking are statistical models and tools for quantifying how the response to a stimuli combination relates to the ensemble of activities evoked when each stimulus is presented independently. We seek solutions under an admixture theory that a single neurons response to multiple stimuli is a dynamically weighted average of its responses to individual items. We evaluated single unit activity in an auditory coding "bottleneck", the inferior colliculus, while monkeys reported the location(s) of one or two simultaneous sounds. Time-domain analyses of recorded spike trains were performed by assuming them to be realizations of inhomogeneous Poisson processes. Admixture behavior was modeled by using transformed Gaussian processes to capture dynamical averaging of single sound response rates. A Markov chain Monte Carlo algorithm was devised to perform Bayesian estimation of idiosyncratic trial level behavior as well as persistent cell level properties.

EO137 Room B4302 OPTIMALITY FOR INSURANCE RISK MODELS Chair: KC Yuen

E0323: **Mean-variance asset-liability management with affine diffusion factor process and a reinsurance option**
Presenter: **Zhongyang Sun**, Sun Yat-sen University, China
An optimal asset-liability management problem for an insurer under the mean-variance criterion is considered. The value of liabilities is described by a geometric Brownian motion while the insurer's risk process is modeled by a general jump process generated by a marked point process. The financial market consists of one risk-free asset and n risky assets with the market price of risk relying on an affine diffusion factor process. By transferring a proportion of insurance risk to a reinsurer and investing the surplus into the financial market, the insurer aims to maximize the expected terminal net wealth and, at the same time, minimize the variance of the terminal net wealth. By using a backward stochastic differential equation (BSDE) approach, closed-form expressions for the efficient frontier and efficient strategy are derived.

E0402: **Optimal excess-of-loss reinsurance and dividend under thinning dependence**
Presenter: **Mi Chen**, Fujian Normal University, China
Co-authors: KC Yuen, Wenyuan Wang
The problem of optimal dividends and reinsurance in a risk model with thinning-dependence structure is studied. Transaction costs and taxes are required when dividends occur. The expected value premium principle is adopted and non-cheap reinsurance is considered. It shows that an excess-of-loss reinsurance strategy is an optimal reinsurance form. Closed-form expressions for the value function and the corresponding optimal excess-of-loss reinsurance strategy are derived. Some numerical examples are presented.

E0390: **Optimal risk control with both fixed and proprotional transaction costs**
Presenter: **Ming Zhou**, Central University of Finance and Economics, China
Co-authors: KC Yuen
A large insurance company is considered whose cumulative cash flow process is described by a drifted Brownian motion. The decision maker of the company has an option to purchase proportional reinsurance at a point of time to minimize the ruin probability and maximize the expected present value of dividend payments up to the time of ruin. In view of the expenses like consultant commission in practice, it is assumed that a fixed transaction cost occurs at the beginning of a reinsurance treaty, and that the reinsurance is irreversible. For this mixed problem of optimal stopping time and stochastic control, we are able to derive the optimal time to purchase the reinsurance, the optimal retained proportion, the optimal dividend barrier, and the value function. The optimal solution shows that reinsurance is valueless to the firm value when the fixed cost is large, and comes into play when the fixed cost is moderate. We also carry out some numerical studies to assess the impact of the fixed cost on the value function and the optimal policies.

E0723: **Annuitization and asset allocation under exponential utility**
Presenter: **Xiaoqing Liang**, Hebei University of Technology, China
Co-authors: Virginia Young
The optimal investment, consumption, and annuitization strategies for a retiree who wishes to maximize her expected discounted utility of lifetime consumption is found. We assume that the retirees preferences exhibit constant absolute risk aversion (CARA), that is, the retirees utility function is exponential. The retiree invests in a financial market with one riskless and one risky asset, the so-called Black Scholes market. Moreover, the retiree may purchase single-premium immediate life annuity income that is payable continuously, and she may purchase this life annuity income at any time and for any amount, subject to the limit of her available wealth. Because maximizing exponential utility generally does not prevent wealth from dropping below 0, we restrict the investment, consumption, and annuitization strategies so that wealth remains non-negative. We solve the

optimization problem via stochastic control and obtain semi-explicit solutions by using the Legendre dual. We prove that the optimal annuitization strategy is a barrier strategy. We also provide some numerical examples to illustrate our results and to analyze their sensitivity to the parameters.

E0374: **Optimal proportional reinsurance to minimize the probability of drawdown under thinning-dependence structure**
Presenter: **KC Yuen**, HKU, China

The optimal proportional reinsurance problem is considered in a risk model with the thinning-dependence structure, and the criterion is to minimize the probability that the value of the surplus process drops below some fixed proportion of its maximum value to date which is known as the probability of drawdown. The thinning dependence assumes that stochastic sources related to claim occurrence are classified into different groups, and that each group may cause a claim in each insurance class with a certain probability. By the technique of stochastic control theory and the corresponding Hamilton-Jacobi-Bellman equation, the optimal reinsurance strategy and the corresponding minimum probability of drawdown are derived not only for the expected value principle but also for the variance premium principle. Finally, some numerical examples are presented to show the impact of model parameters on the optimal results.

EO059 Room G4302 RECENT ADVANCES IN TIME SERIES AND SPATIAL ECONOMETRICS AND STATISTICS Chair: Zudi Lu

E0198: **Time-varying graphs by locally stationary Hawkes processes**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Yu Izumisawa, Junichi Hirukawa, Taiga Uno

Hawkes Graphs have been recently introduced to grasp the branching structure of multivariate stationary Hawkes processes. However, existing procedure cannot describe the time structural changes since stationary Hawkes processes are a class of stationary processes. We consider a multivariate locally stationary Hawkes (lsHawkes) process, which is a natural extension of a previously studied univariate lsHawkes process. We first consider an approximation of the lsHawkes process as a time-varying integer-valued autoregressive (tvINAR) process. Then, we propose an estimation procedure for the time varying parameters based on local least-squares method. Finally, we propose time-varying Hawkes graphs (tvHawkes graphs) based on the estimated parameters.

E0163: **Hybrid quantile regression estimation for time series models with conditional heteroscedasticity**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong

Estimating conditional quantiles of financial time series is essential for risk management and many other financial applications. For time series models with conditional heteroscedasticity, although it is the generalized autoregressive conditional heteroscedastic (GARCH) model that has the greatest popularity, so far, only a variant of the GARCH model, the so-called linear GARCH model, has been feasible for quantile regression. An easy-to-implement hybrid quantile regression estimation procedure for the GARCH model is proposed, where we overcome the intractability due to the square-root form of the corresponding conditional quantile function by a simple transformation. The proposed method takes advantage of the efficiency of the GARCH model in modeling the volatility globally as well as the flexibility of the quantile regression in fitting quantiles at a specific level. The asymptotic distribution of the proposed estimator is derived and is approximated by a novel mixed bootstrapping procedure. A Portmanteau test is further constructed to check the adequacy of fitted conditional quantiles. The finite-sample performance of the proposed method is examined by simulation studies, and its advantages over existing methods are illustrated by an empirical application to Value-at-Risk forecasting.

E0370: **Robust factor model with partially explained covariates**
Presenter: **Yuan Ke**, Penn State University, United States

Factor models are studied when the latent factors can be partially explained by observed covariates. With those covariates, both the factors and loadings are identifiable up to a rotation matrix even only with a sufficiently large finite dimensions. To incorporate the explanatory power of these covariates, we propose a smoothed principal component analysis (PCA): 1 regress the data onto the observed covariates, and 2 take the principal components of the fitted data to estimate the loadings and factors. We show that both the estimated factors and loadings can be estimated with improved rates of convergence compared to the benchmark method. The degree of improvement depends on the strength of the signals, representing the explanatory power of the covariates on the factors. The proposed estimator is robust to possibly heavy-tailed distributions, which are encountered in many high-dimensional applications for factor analysis. Empirically, our method leads to a substantial improvement on the out-of-sample forecast on the US bond excess return data.

E0559: **Semiparametric regularisation and estimation for partially nonlinear spatio-temporal regression models**
Presenter: **Zudi Lu**, University of Southampton, United Kingdom
Co-authors: Dawlah Alsulami, Zhenyu Jiang

Semiparametric modelling of spatio-temporal data has received increasing attention owing to its flexibility in uncovering potential nonlinear impact of a covariate on the response variable of spatio-temporal nature. For example, to model the possible nonlinear relationship between Consumer Price Index (CPI) and Housing Price Index (HPI) in United States (US), accounting for the spatio-temporal lag effects of neighbouring states would give more accurate estimation and prediction. In doing this, a fundamental difficulty is how to flexibly account for the spatio-temporal neighbouring lag effects. We propose two data-driven schemes to address solving this difficulty by extending the semiparametric regularisation methodology to simultaneously estimate and select the important spatio-temporal neighbouring lag variables for semiparametric modelling of spatial time series data. We allow the data are non-stationary over all spatial locations that are irregularly positioned on the earth surface (but stationary along time). New estimation procedures are developed with an improved family of data-driven semiparametric spatio-temporal regression models for both estimation and selection. The real data application demonstrates the proposed new models can significantly improve spatio-temporal prediction than the existing semiparametric spatio-temporal modelling.

E0575: **Least squares estimation for nonlinear regression models with heteroscedasticity**
Presenter: **Qiying Wang**, University of Sydney, Australia

Asymptotic theory is developed for general nonlinear regression models, establishing a new framework on least squares estimation that is easy to apply for various nonlinear regression models with heteroscedasticity. An application of the framework to nonlinear regression models with nonstationarity and heteroscedasticity is explored. Accompanying with these main results, a maximum inequality for a class of martingales is provided and some new results are established on convergence to a local time and convergence to a mixture of normal distributions.

EO155 Room G4701 MIXED-EFFECTS MODELS AND STATISTICAL MODELING FOR COMPLEX DATA Chair: Dalei Yu

E0290: **Principal Hessian directions for mixture multivariate skew elliptical distributions**
Presenter: **Fei Chen**, Yunnan University of Finance and Economics, China
Co-authors: Lei Shi, Lixing Zhu, Xuehu Zhu

As a nice application of Stein's lemma, principal Hessian directions (pHd) using Hessian matrix is a moment-based method and becomes a promising methodology in sufficient dimension reduction because of its easy implementation. However, it requires strong conditions on the distribution of the predictors, which is very close to the normality assumption. Out of curiosity for its theoretical development and caution for its practical use, we investigate whether and how pHd is applicable when the distribution is mixture multivariate skew elliptical (MMSE) distribution and if not, how to derive a generalized pHd. Further, we propose two new estimation algorithms for the generalized pHd. Numerical studies are conducted to examine its performance in finite sample cases. The theoretical results also serves as a reminder for researchers and users to pay more attention to the theoretical conditions as pHd critically relies on them.

E0285: **A competing risk model with bivariate random effects for clustered survival data**
Presenter: **Xin Lai**, Xian Jiaotong University, China
Co-authors: Kelvin Yau, Liu LIU

Competing risks are often observed in clinical trial studies. As exemplified in two data sets, the bone marrow transplantation study for leukemia patients and the primary biliary cirrhosis study, patients could experience two competing events which may be correlated due to shared unobservable factors within the same cluster. With the presence of random hospital/cluster effects, a cause-specific hazard model with bivariate random effects is proposed to analyze clustered survival data with two competing events. This model extends earlier work by allowing random effects in two hazard function parts to follow a bivariate normal distribution, which gives a generalized model with a correlation parameter governing the relationship between two events due to the hospital/cluster effects. By adopting the GLMM formulation, random effects are incorporated in the model via the linear predictor terms. Estimation of parameters is achieved via an iterative algorithm. A simulation study is conducted to assess the performance of the estimators, under the proposed numerical estimation scheme. Application to the two sets of data illustrates the usefulness of the proposed model.

E0751: **Estimation and confidence regions in models with orthogonal block structure**
Presenter: **Sandra Ferreira**, University of Beira Interior, Covilha, Portugal
Co-authors: Dario Ferreira, Celia Nunes, Joao Mexia

The first purpose is to provide an overview of the algebraic structure of the family of models with orthogonal block structure, OBS. Then we will discuss optimal properties for the estimators in this class of models. An example about the construction of confidence regions for the parameters of these model is presented.

E0752: **Bayesian cumulative logit random effects models for longitudinal ordinal data**
Presenter: **Jiyeong Kim**, Sungkyunkwan University, Korea, South
Co-authors: Keunbaik Lee

In analysis of longitudinal categorical data, generalized linear mixed models (GLMMs) are typically used. The random effects covariance matrix in the GLMMs explains both subject-specific and serial correlation of repeated outcomes. However, estimation of the covariance matrix is not easy because of high dimensionality and positive-definiteness of the estimated one. In addition, the structure of the covariance matrix can be heteroscedastic. To solve these constraints, Cholesky-type decompositions of the covariance matrix have been proposed (modified Cholesky, moving-average Cholesky and autoregressive moving-average Cholesky decomposition). We analyze longitudinal ordinal data using one of GLMMs, cumulative logit random effects models (CLREMs) with autoregressive moving-average random effects covariance matrix. In addition, various Cholesky-type decompositions are compared to our methods. The method are illustrated by a lung cancer data set.

E0293: **On the asymptotic properties and information criteria for misspecified generalized linear mixed models**
Presenter: **Dalei Yu**, Yunnan University of Finance and Economics, China
Co-authors: Xinyu Zhang, Kelvin Yau

The problem of misspecification poses challenges in model selection. We study the asymptotic properties of estimators (including the estimators of fixed effects, variance component parameters and predictors of random effects) for generalized linear mixed models with misspecification under the framework of conditional Kullback-Leibler divergence. A conditional generalized information criterion is introduced, and a model selection procedure is proposed by minimizing the criterion. We prove that the proposed model selection procedure is asymptotically loss efficient when all the candidate models are misspecified. We also investigate the model selection consistency of the proposed model selection criterion. Numerical studies confirm the effectiveness of the proposed criterion.

EO042 Room LT-11 RECENT ADVANCES IN ECONOMETRIC THEORY AND METHODS Chair: Kaiji Motegi

E0606: **Testing identifying assumptions in a fuzzy regression discontinuity design**
Presenter: **Toru Kitagawa**, University College London, United Kingdom
Co-authors: Yoichi Arai, Yu-Chin Hsu, Ismael Mourifie, Yuanyuan Wan

A new specification test is proposed for the validity of fuzzy regression discontinuity designs (FRD-validity). We derive a new set of testable implications for FRD-validity which is characterized by a set of inequality restrictions on the joint distribution of observed outcomes and treatment status at the cut-off. We show that it exploits all the information in data useful for screening out violation of FRD-validity. Our approach differs from and complements the existing approaches that test continuity of the distributions of running variables and baseline covariates at the cut-off since ours focuses on the distribution of observed outcome and treatment status. We show that the proposed test controls size in large sample uniformly over a large class of distributions, is consistent against all fixed alternatives, and has non-trivial power against local alternatives. Two empirical applications illustrate uses of our test.

E0627: **Model checks for nonlinear cointegrating regression**
Presenter: **Ke Zhu**, University of Hong Kong, Hong Kong

Using the marked empirical processes, a test of parametric specification in a nonlinear cointegrating regression model is developed. Unlike the kernel-smoothed U-statistic, the new test statistic avoids the use of bandwidth, which has some advantages for practitioners. Simulations and a real data example show that our new test has a good finite sample performance. Other contributions are to provide a rigorous proof on weak convergence for a class of martingales and construct a simulated estimator of the limiting null distribution, which are interesting in their own rights.

E0254: **Robust statistical inference for time series regression model by self-normalized subsampling method**
Presenter: **Fumiya Akashi**, Waseda University, Japan
Co-authors: Shuyang Bai, Murad Taqqu
Robust statistical inference for possibly long-memory and/or heavy-tailed processes is considered. In the context of time series analysis, we often observe heavy-tailed and long-range dependent data in variety of fields. To model such data suitably, we consider a linear regression model with dependent covariate and error processes. When the model has heavy-tails or long-memory, it is well known that fundamental statistics (e.g., sample mean) converge to involved distributions and the rate of convergence of the statistic contains unknown tail-index and Hurst index. To overcome such difficulties, we propose the self-normalized statistic and subsampling procedures. As a result, we construct a confidence region for the regression parameter of the model without any prior estimation of nuisance parameters. Some simulation experiments illustrate the finite sample performance of the proposed method.

E0730: **A simple and efficient estimation method for models with nonignorable missing data**
Presenter: **Zheng Zhang**, Renmin University of China, China
Co-authors: Chunrong Ai, Oliver Linton
A simple and efficient estimation procedure is proposed for the model with non-ignorable missing data recently studied. The previous semiparametrically efficient estimator requires explicit nonparametric estimation, and thus, it suffers from the curse of dimensionality and requires a bandwidth selection. We propose an estimation method based on the Generalized Method of Moments (hereafter GMM). Our method is consistent and asymptotically normal regardless of the number of moments chosen. Furthermore, if the number of moments increases appropriately our estimator can achieve the semiparametric efficiency bound derived previously, but under weaker regularity conditions. Moreover, our proposed estimator and its consistent covariance matrix are easily computed with the widely available GMM package. We propose two databased methods for selection of the number of moments. A small scale simulation study reveals that the proposed estimation indeed outperforms the existing alternatives in finite samples.

E0210: **Calibration estimation for semiparametric copula models under missing data**
Presenter: **Kaiji Motegi**, Kobe University, Japan
Co-authors: Zheng Zhang, Shigeyuki Hamori
The estimation of semiparametric copula models under the presence of missing data is investigated. Our models comprise nonparametric marginal distributions and parametric copula functions. The two-step pseudo-likelihood method is infeasible when there exist missing data. Inspired in a recent work, we propose a class of calibration estimators for both marginal distributions and the parameters of interest without imposing additional models on the missing mechanism. We establish consistency and asymptotic normality for our estimators of copula parameters. We also present a natural procedure for consistently estimating the asymptotic variance of our estimators.

EO186 Room LT-13 FINANCIAL ECONOMETRICS WITH HIGH FREQUENCY DATA Chair: Binyan Jiang

E0270: **Factor correlation matrix modelling of large-dimensional portfolio with high-frequency data**
Presenter: **Yingjie Dong**, University of International Business and Economics, China
Co-authors: Yiu-Kuen Tse
A factor correlation matrix approach is proposed to model high-dimensional realized covariance matrix using high-frequency data. We assume the high-dimensional daily realized correlation matrix is driven by a low-dimensional latent process, which is modelled using the principal component method. We adopt the vech representation for the low-dimensional latent process over time. In addition, the return variances are estimated by imposing a long memory structure on the realized volatilities. We conduct Monte Carlo studies to compare the finite sample performance of different methods of estimating the high-dimensional covariances. Our new method is found to perform better by reporting smaller estimation errors. In addition, our empirical studies show that our method provides lower variance in selecting minimum-variance portfolios.

E0273: **Testing equality of principle components in factor models**
Presenter: **Ningning Xia**, Shanghai University of Finance and Economics, China
Co-authors: Jianqing Fan, Yingying Li, Xinghua Zheng
A test is developed for structural breaks in factor models. We focus on the null hypothesis that the size and direction of factor are constant over time. Because structural change can happen to either the factor loadings or the dynamic process of factors or both, we can not only focus on the change of factor loadings. Based on the fact that the presence of a structural change in factor model yields a size or direction change of principle components obtained from the sample covariance matrix, we test the equality of sample principle components and compare the pre and post break subsample covariance matrix. Our test is consistent under the alternative hypothesis in which a fraction of or all factors have structural changes. The simulation studies show that our test has good finite-sample size and power.

E0631: **Penalized interaction estimation for ultrahigh dimensional quadratic regression**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong
Quadratic regression goes beyond linear model by simultaneously including main effects and interactions between the covariates. The problem of interaction estimation in high dimensional quadratic regression has received extensive attention in the past decade. We introduce a novel method which allows us to estimate the main effects and interactions separately. Unlike existing methods for ultrahigh dimensional quadratic regressions, our proposal does not require the widely used heredity assumption. In addition, our proposed estimates have explicit formulas and obey the invariance principle at the population level. We estimate the interactions of matrix form under penalized convex loss function. The resulting estimates are shown to be consistent even when the covariate dimension is an exponential order of the sample size. We develop an efficient ADMM algorithm to implement the penalized estimation. This ADMM algorithm fully explores the cheap computational cost of matrix multiplication and hence is much more efficient than existing penalized methods under heredity constraints. We demonstrate the promising performance of our proposal through extensive numerical studies.

E0794: **Testing if the market microstructure noise is a function of the limit order book**
Presenter: **Simon Clinet**, Keio University, Japan
Co-authors: Yoann Potiron
The aim is to build tests for the presence of error in a model where the market microstructure noise is a known parametric function of the limit order book. The tests compare two novel and distinct quasi-maximum likelihood estimators of volatility, where the related model includes an additive error in the market microstructure noise or not. The limit theory is investigated in a general nonparametric framework. When there is no error in the model, following a common procedure, we provide a consistent estimator of the efficient price based on maximum likelihood estimation of the parameter. Furthermore, we show that realized volatility remains efficient when performed on the estimated price rather than on the efficient price.

E0779: **Autoregressive conditional duration model with time-varying parameters for discrete trade durations**
Presenter: **Petra Tomanova**, University of Economics, Prague, Czech Republic
Autoregressive conditional duration (ACD) models are useful for modelling the time between events, in particular the time between trading of stocks, which is known as trade durations. However, trade durations exhibit certain issues such as excessive zeros, discrete nature, overdispersion and intraday seasonality. We propose ACD model under the generalized autoregressive score framework allowing parameters to vary over time and capturing the dynamics of time-varying parameters by the autoregressive term and the scaled score of the conditional observation density. To deal with the discreteness of the data, we consider the trade durations as non-negative integer variables. This kind of variables is commonly analyzed using count data models based on specific underlying distribution, most notably the Poisson distribution. However, the feature that its expected value is equal to its variance is too strict in many applications as count data often exhibit overdispersion. To overcome this limitation we assume trade durations to follow the negative binomial distribution and we extend it to capture excessive zeros using the zero-inflated model. Simulation study and empirical analysis using high frequency data from NYSE TAQ database illustrate that the proposed model outperforms continuous ACD models.

EO251 Room LT-14 MACHINE LEARNING THEORY Chair: Andreas Christmann

E0214: **Universality of deep CNNs and distributed learning**
Presenter: **Ding-Xuan Zhou**, City University of Hong Kong, Hong Kong
The aim is to show that deep convolutional neural networks (CNNs) with the rectified linear unit activation function without any fully-connected layers are universal as the level tends to infinity. We shall also discuss some problems related to distributed learning. Our approach is based on machine learning and approximation theory.

E0211: **Total stability of support vector machines**
Presenter: **Daohong Xiang**, Zhejiang Normal University, China
Co-authors: Andreas Christmann, Ding-Xuan Zhou
Some total stability results for SVMs are established, which show that SVMs based on kernels are even stable, if the full triple (P, λ, k) consisting of the underlying probability measure P, the regularization parameter λ, and the kernel k changes slightly.

E0218: **Statistical learning for modal regression**
Presenter: **Yunlong Feng**, The State University of New York at Albany, United States
Co-authors: Jun Fan, Johan Suykens
The modal regression problem will be discussed from a statistical learning point of view. It will be shown that modal regression can be approached by means of empirical risk minimization techniques. A framework for analyzing and implementing modal regression within the statistical learning context will be developed. Theoretical results concerning the generalization ability and approximation ability of modal regression estimators will be provided. Connections and differences of the proposed modal regression method with existing ones will also be illustrated. Numerical examples will be given to show the effectiveness of the newly proposed modal regression approach.

E0248: **Lepskii principle in supervised learning**
Presenter: **Nicole Muecke**, IIT@MIT-USA, Italy
A statistical supervised learning problem is investigated under random design. In particular, we analyze reproducing kernel-based estimators arising from a fairly large class of spectral regularization methods. We derive a fully adaptive data driven estimator via a slightly modified version of Lepskiis principle, giving oracle properties both in reproducing kernel norm and prediction norm.

E0429: **Robustness and stability of kernel based learning**
Presenter: **Andreas Christmann**, University of Bayreuth, Germany
Statistical machine learning and methods for big data are very important topics in current research. The fourth dimension of big data is often called veracity, i.e. handling data in doubt. This concept is similar to the notion of statistical robustness. Certain questions of universal consistency, statistical robustness, and stability will be addressed.

EO053 Room LT-17 MODELLING COMPLEX TIME SERIES: ESTIMATION AND FORECASTING Chair: Hsein Kew

E0345: **An iterative approach for model selection in single-index varying coefficient models**
Presenter: **Efang Kong**, University of Electronic Science and Technology of China, China
The penalised least squares estimation based model selection has many advantages over the traditional ones. Much literature has been devoted to this area in recent years. The single-index varying coefficient models (SIVC) have proved to be a class of very useful models in data analysis. Model selection in such class is important but challenging due to the complicated structure of SIVC. We take on this challenge and develop an iterative approach for model selection in SIVC. The proposed approach is easy to implement as there is a closed form for estimators obtained in each step. Asymptotic properties of the proposed iterative approach are also established, which provide theoretical justification for the proposed approach. Comprehensive simulation studies show that the proposed iterative approach works very well even with modest sample size. Finally, we apply the SIVC and the proposed model selection method to an environmental set from Hong Kong and the Boston housing dataset from Boston, which lead to some interesting findings.

E0423: **On the distribution of US banks size over time**
Presenter: **Yong Song**, University of Melbourne, Australia
Co-authors: John Maheu
Bank size distribution is an important observable measure to the central monetary authority. In a highly concentrated banking sector, a few large banks may form a certain oligopoly structure and incur economic efficiency loss. On the other hand, an over competitive banking industry could be associated with financial instability, which imposes a systemic risk to the health of the whole economy. In addition to existing parametric change-point models, we propose a novel Bayesian nonparametric distribution change framework to monitor the evolution of the back size distribution over time.

E0473: **Adaptive estimation of semi-parametric partially linear predictive regression under heteroskedasticity**
Presenter: **Hsein Kew**, Monash University, Australia
Co-authors: Jiti Gao, Yundong Tu
Adaptive estimation is considered in semiparametric partially linear predictive regression models with unconditional heteroscedasticity of an unknown form. We develop an adaptive semiparametric estimator weighted by a non-parametric variance estimator. The adaptive estimator is shown to deliver potentially large asymptotic efficiency gains over the conventional unweighted estimator. Monte Carlo simulations confirm this

theoretical result. We implement the proposed estimation method by studying the in-sample predictability of US future stock returns using the commonly used financial variables as regressors.

E0562: **Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series**
Presenter: **Jia Chen**, University of York, United Kingdom
Two semiparametric model averaging schemes are proposed for nonlinear dynamic time series regression models with a very large number of covariates. The objective is to obtain accurate estimates and forecasts of time series nonparametrically. In the first scheme we use a Kernel Sure Independence Screening technique to screen out insignificant regressors; we then use a semiparametric penalized method of Model Averaging MArginal Regression for the regressors that have survived the screening procedure, to further select regressors that have significant effects on estimating the multivariate regression function and predicting the future values of the response variable. In the second scheme, we impose an approximate factor modelling structure on the ultra-high dimensional exogenous regressors and use the principal component analysis to estimate the latent common factors; we then apply the penalised Model Averaging MArginal Regression method to select significant common factors and lags of the response variable. In each of the two schemes, we construct the optimal combination of the significant marginal regression and auto-regression functions. Asymptotic and numerical studies of the proposed methods are provided.

E0582: **Eigen portfolio selection: A robust approach to Sharpe ratio maximization**
Presenter: **Danqiao Guo**, University of Waterloo, Canada
Co-authors: Phelim Boyle, Chengguo Weng, Tony Wirjanto
It is shown how to pick optimal portfolios by modulating the impact of estimation risk in large covariance matrices. The portfolios are selected to maximize their Sharpe ratios. Each eigenvector of the covariance matrix corresponds to a maximum Sharpe ratio (MSR) portfolio for a different set of expected returns. Assuming that the portfolio manager has views on the future expected returns, a portfolio consistent with her views can be approximated by the first few eigenvectors of the sample covariance matrix. Since the estimation error in a large sample 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. To seek a balance between the approximation error and the estimation error, we set a tolerance limit for the former and make best efforts to control the latter. We further introduce a more general spectral selection method, which uses non-consecutive eigenvectors to approximate the expected excess returns. According to simulation and real-data studies, the advantage of the spectral selection method becomes apparent when the number of assets is large compared with the sample size.

EO040 Room P4701 RECENT SEMI/NONPARAMETRIC STATISTICAL DEVELOPMENTS AND THEIR APPLICATIONS Chair: Heng Lian

E0231: **Variable selection via penalized GEE for a marginal survival model**
Presenter: **Yi Niu**, Dalian University of Technology, China
Clustered and multivariate survival times, such as times to recurrent events, often arise in biomedical and health research, and marginal survival models are often used to model such data. When there are a large number of predictors available, variable selection is always an important issue when modeling such data with a marginal survival model. We consider a marginal Cox's proportional hazards model. Under the sparsity assumption, we propose a penalized generalized estimating equations approach to select important variables and to estimate regression coefficients simultaneously in the marginal model. The proposed method explicitly models the correlation structure within clusters or correlated variables by using a prespecified working correlation matrix. The asymptotic properties of the estimators from the penalized generalized estimating equations are established and the number of candidate covariates is allowed to increase in the same order as the number of clusters does. We evaluate the performance of the proposed method through a simulation study and demonstrate its application using two real datasets.

E0329: **Stein discrepancy methods for robust estimation**
Presenter: **Emre Barut**, George Washington University, United States
All statistical procedures highly depend on the modeling assumptions and how close these assumptions are to reality. This dependence is critical: Even the slightest deviation from assumptions can cause major instabilities during statistical estimation. In order to mitigate issues arising from model mismatch, numerous methods have been developed in the area of robust statistics. However, these approaches are aimed at specific problems, such as heavy tailed or correlated errors. The lack of a holistic framework in robust regression results in a major problem for the data practitioner. That is, in order to build a robust statistical model, possible issues in the data have to be found and understood before conducting the analysis. In addition, the practitioner needs to have an understanding of which robust models can be applied in which situations. We propose a new framework for parameter estimation, which is given as the empirical minimizer of a second order U-statistic. When estimating parameters in the exponential family, the estimate can be obtained by solving a quadratic convex problem. For parameter estimation, our approach significantly improves upon MLE when outliers are present, or when the model is misspecified. Furthermore, we show how the new estimator can be used to efficiently fit to distributions with unknown normalizing constants. Extensions of our method for regression problems and implications for statistical modeling are discussed.

E0521: **Multivariate tests in high dimensions and unstructured dependence**
Presenter: **Solomon Harrar**, University of Kentucky, United States
Co-authors: Xiaoli Kong
Recent results for comparison of high-dimensional mean vectors make assumptions that requires the dependence between the variables to be weak. This requirement fails to be satisfied, for example, by elliptically contoured distributions. We relax the dependence conditions that seem to be the standard assumption in high-dimensional asymptotic tests. With the relaxed condition, the scope of applicability of the the results broadens. In particular, strong mixing type of dependence and applications for rank-based comparison of groups are covered. For the rank-based methods, hypotheses are formulated in terms of meaningful and easy to interpret measures of effects for the nonparametric methods. This formulation accommodate data in binary, discrete, ordinal and continuous scales seamlessly. The problem is setup in a general and flexible form that extension of the results to general factorial design, including repeated measures, are formally illustrated. Simulation studies are used to evaluate the numerical performance of the results in practical scenarios. Data from Electroencephalograph (EEG) experiment is analyzed to illustrate the application of the results.

E0455: **Analysis of longitudinal data anchored by interval censored events**
Presenter: **Ying Zhang**, Indiana University, United States
Co-authors: Chenghao Chu, Wanzhu Tu
In many longitudinal studies, outcomes are assessed on time scales anchored by certain clinical events. When the anchoring events are unobserved, the study timeline becomes undefined, and the traditional longitudinal analysis loses its temporal reference. We consider the analytical situations where the anchoring events are interval censored. We show that by expressing the regression parameter estimators as stochastic functionals of a plug-in estimate of the unknown anchoring event distribution, the standard longitudinal models can be modified and extended to accommodate the less well defined time scale. This extension enhances the existing tools for longitudinal data analysis. Under mild regularity conditions, we

show that for a broad class of models including the frequently used generalized linear mixed-effects models, the functional parameter estimates are consistent and asymptotically normally distributed with root-n convergence rate. To implement, we developed a hybrid computational procedure combining the strength of the Fisher's scoring method and the EM algorithm. We conducted a simulation study to validate the asymptotic properties, and to examine the finite sample performance of the proposed method. A real data analysis was used to illustrate the proposed method.

E0755: **Sufficient variable selection using independence measures for continuous responses**
Presenter: **Baoying Yang**, Southwest Jiaotong University, China
Two sufficient variable selection procedures are proposed: one stage and two stage approaches using independence measures for continuous responses. Although any independence measure can be used, we use distance correlation and Hilbert Schmidt Independence Criterion correlation to illustrate the two procedures. By comparing with some existing marginal screening methods, we show the advantages of the proposed procedures through the simulation studies and a real data analysis. Our procedures are model-free and robust against model mis-specification, especially useful, when the active predictors are marginally independent of the response.

EO162 Room P4703 MODERN STATISTICAL METHODS FOR QUALITY ENGINEERING Chair: Mei Han

E0162: **Optimal robust and tolerance design for computer experiments with mixture proportion inputs**
Presenter: **Mei Han**, City University of Hong Kong, Hong Kong
Computer experiments often have inputs that are proportions/fractions of components in a mixture. In these mixture computer experiments, it can be of interest to perform robust and tolerance design on the mixture proportions since the proportions are subjected to noise variations. Traditionally, manufacturing of mixture products is controlled via interval tolerances for mixture amounts. An optimal tolerance region for proportions, which gives optimal quality cost among all possible tolerance regions for mixture proportions with the same acceptance probability, is proposed for integrated parameter and tolerance design (IPTD) in mixture computer experiments. Real examples are given to demonstrate the improvements that can be achieved with the optimal tolerance region.

E0179: **Effective model calibration via sensible variable identification and adjustment**
Presenter: **Yan Wang**, City university of Hong Kong, Hong Kong
Co-authors: Xiaowei Yue, Rui Tuo, Jeffrey H Hunt
Computer models often fail to fit the physical system perfectly because of all kinds of assumptions for the purpose of mathematical tractability. Estimation of model parameters of computer simulations, also known as calibration, is an important topic in many engineering applications. We consider the calibration of computer model parameters with the help of engineering design knowledge. We introduce the concept of sensible (calibration) variables. Sensible variables are model parameters which are sensitive in the engineering modeling, and whose optimal values are different from the pre-specified design values. We propose an effective calibration method to identify and adjust the sensible variables with limited physical experimental data. A numerical study and a composite fuselage simulation example show the effectiveness and efficiency of the proposed method.

E0217: **Online Bayesian optimization design-based closed-loop control with model parameter uncertainty and data quality**
Presenter: **Linhan Ouyang**, Nanjing University of Aeronautics and Astronautics, China
Response surface-based design optimization has been commonly used to seek the optimal input settings in processes or products design problems. Focused on statistical modeling and numerical optimization strategies, most researchers typically assume that there is no model parameter uncertainty in modeling process or the data quality performs well in the online update process. However, if the estimated model parameter varies from the true one or the online observations contain substantial error in quality improvement, the resulting solution may be quite far from the optimal. A Bayesian modeling approach is proposed to closed-loop online optimization design that accounts for model parameter uncertainty and data quality in micro-manufacturing processes. The uniqueness of the proposed approach can provide the information of how and when to update the settings of the design variables based on the online observations. Therefore, it can avoid the danger of over-updating the process if online design approaches are used in quality improvement. The effectiveness of the approach is illustrated with simulation experiments and a micro-milling process. The comparison results demonstrate that the proposed approach with consideration of model parameter uncertainty and data quality can achieve better process performance than conventional design approaches, since it can make corrective adjustments by updating the model parameters or the target value of each run.

E0222: **Quality design for laser micro-manufacturing processes using Bayesian modeling and optimization methods**
Presenter: **Jianjun Wang**, Nanjing University of Science and Technology, China
The micro-manufacturing process usually has some typical features such as multiple noise factors, large variations, high manufacturing costs and poor repeatability. In view of high uncertainties for quality design problems in the micro-manufacturing process, a unified framework of Bayesian modeling and optimization is proposed to improve product quality and reduce manufacturing cost. First of all, Bayesian modeling methods are utilized to develop the relationship models between input factors (e.g., laser average power, pulse frequency and cutting speed) and output quality characteristics (e.g., hole diameter and roundness) in the laser micro-drilling process. Then, the simulated responses which reflect a real laser micro-drilling process are obtained by using a Gibbs sampling procedure. Furthermore, the cost structures for mixed multiple quality characteristics is analyzed and then the rejection cost (i.e., rework cost and scrap cost) function is constructed by using the simulated response values. Finally, the optimum economic parameter settings of laser micro-drilling process can be obtained by optimizing the proposed cost function with a hybrid genetic algorithm. The results show the proposed approach can significantly improve the product quality and reduce the rejection cost in the micro-drilling process.

E0327: **On-line monitoring data quality of high-dimensional data streams**
Presenter: **Zhonghua Li**, Nankai University, China
In recent years, effective monitoring of data quality has increasingly attracted attention of researchers in the area of statistical quality control. Among the relevant research on this topic, none used multivariate methods to control the multidimensional data quality process, but instead relied on multiple univariate control charts. Based on a novel one-sided multivariate exponentially weighted moving average (MEWMA) chart, a conditional false discovery rate (FDR) adjusted scheme will be introduced to on-line monitor the data quality of high-dimensional data streams. With thousands of input data streams, the average run length loses its usefulness because one will likely have out-of-control signals at each time period. Hence, the FDR and power are chosen as two criteria used for the performance comparison. Numerical results show that the proposed MEWMA scheme has both less conservative FDR and high average power.

EO147 Room P4704 SEMI- AND NONPARAMETRIC INFERENCE IN SURVIVAL ANALYSIS AND RELIABILITY Chair: Eric Beutner

E0348: **Locally weighted regression quantiles with competing risks**
Presenter: **Sangbum Choi**, Korea University, Korea, South
Flexible estimation and inference procedures are considered for competing risks quantile regression that not only provide meaningful interpretations by using cumulative incidence quantiles, but also extend the conventional accelerated failure time model by relaxing some of the stringent model assumptions such as global linearity and unconditional independence. The locally weighed technique for censored quantile regressions is extended to the competing risks setting. The proposed procedure permits the fast and accurate computation of quantile regression parameter estimates and standard variances by using conventional numerical methods. Numerical studies show that the proposed estimators perform well and the resulting inference is reliable in practical settings. The method is finally applied to data from a soft tissue sarcoma study.

E0367: **Likelihood-based inference for latent failure time models with competing risks under the generalized FGM copula**
Presenter: **Jia-Han Shih**, National Central University, Taiwan
Co-authors: Takeshi Emura
Many existing latent failure time models for competing risks do not provide closed form expressions of sub-distribution functions. We suggest a generalized FGM copula model with the Burr III failure time distribution such that the sub-distribution functions have closed form expressions. Under the suggested model, we develop a likelihood-based inference method along with its computational algorithms. Based on the expressions of the sub-distribution functions, we propose two goodness-of-fit tests: (I) The overall model fit and (II) The fit of the generalized FGM copula. Our tests are based on the parametric bootstrap which utilize the parametric, semi-parametric, and non-parametric estimators of sub-distribution functions. Simulations are conducted to examine the performance of the proposed methods. A real data from the reliability analysis of the radio transmitter-receivers are analyzed for illustration.

E0484: **Resampling-based inference for the Mann-Whitney effect for right-censored and tied data**
Presenter: **Dennis Dobler**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Markus Pauly
In a two-sample survival setting with independent survival variables T and R and independent right-censoring, the Mann-Whitney effect $p = P(T > R) + 0.5P(T = R)$ is an intuitive measure for discriminating two survival distributions. Comparing two treatments, the case $p > 0.5$ suggests the superiority of the first. Nonparametric maximum likelihood estimators based on normalized Kaplan-Meier estimators naturally handle tied data, which are omnipresent in practical applications. Studentizations allow for asymptotically accurate inference on p. For small samples, however, coverage probabilities of confidence intervals are considerably enhanced by means of bootstrap and permutation techniques. The latter even yields finitely exact procedures in the situation of exchangeable data. Simulation results support all theoretic properties under various censoring and distribution set-ups.

E0647: **Semiparametric methods for recurrent event times models with application to virtual age models**
Presenter: **Eric Beutner**, Maastricht University, Netherlands
Co-authors: Laurent Bordes, Laurent Doyen
Virtual age models are very useful to analyse recurrent events. Among the strengths of these models is their ability to account for treatment (or intervention) effects after an event occurrence. Despite their flexibility for modeling recurrent events the number of applications is limited. This seems to be a result of the fact that in the semiparametric setting all the existing results assume the virtual age function that describes the treatment (or intervention) effects to be known. This shortcoming can be overcome by considering semiparametric virtual age models with parametrically specified virtual age functions. Yet, fitting such a model is a difficult task. Indeed it has recently been shown that for these models the standard profile likelihood method fails to lead to consistent estimators. We can consider statistical properties of estimators constructed by smoothing the profile log-likelihood function appropriately. We show that our general results derived by smoothing can be applied to most of the relevant virtual age models of the literature. Our approach shows that empirical process techniques may be a worthwhile alternative to martingale methods for studying asymptotic properties of these inference methods.

E0669: **Modelling unbalanced hierarchical survival data using nested Archimedean copula functions**
Presenter: **Roel Braekers**, Hasselt University, Belgium
A copula model for hierarchically nested clustered survival times is introduced in which the different clusters and sub-clusters are possibly unbalanced. Due to the right censoring, we do not fully observe each outcome variable. This, together with the hierarchical structure of the data, makes it difficult to set-up a full likelihood function for a general copula model. To circumvent this problem, we focus hereto on the class of hierarchical nested Archimedean copula functions and use the properties of this copula family to simplify the full likelihood function. For the marginal survival time, we consider a semi-parametric Cox's regression model. Since maximizing the likelihood function for all parameters is computational difficult, we consider a two-stage estimation procedure in which we first estimate the marginal parameters and afterwards estimate the association parameters. As result, we obtained the asymptotic consistency and normality of the association parameters. Next we compare the finite sample behaviour of the different estimators through a simulation study. Furthermore we illustrate this copula model on a practical real life data example.

EP001 Room 4/F University Concourse POSTER SESSION Chair: Maria Elena Fernandez Iglesias

E0310: **Control of 2d-fdr by combining two univariate multiple testing results with application to mass spectral data**
Presenter: **Jaesik Jeong**, Chonnam National University, Korea, South
The mass spectral data feature high dimension with small number of signals (peaks) and many noisy observations. This unique aspect of mass spectral data motivates the problem on testing of many composite null hypotheses simultaneously. We develop new procedures to control the false discovery rate of the simultaneous multiple hypothesis testing of many "bivariate" composite null hypotheses. Two types of (bivariate) composite null hypothesis, the intersection-type and the union-type null, are considered; a different procedure is proposed for each type. The new procedures (for both types of composite null hypotheses) are in two stages. In the first stage, we test simultaneously each "univariate" simple hypotheses of "bivariate" composite hypotheses at the pre-decided false discovery rate, and in the second stage, we combine the marginal univariate test results so that the two-dimensional false discovery rate for the "bivariate" composite null hypotheses is less than alpha, the aimed level. The new procedure provides a closed form decision rule on bivariate test statistics, unlike the existing two-dimensional local false discovery rate (2d-fdr). We numerically compare the performance of our procedure (for the union-type composite null) to the existing 2d-fdr under various settings. We then apply the procedure to the problem of differentiating origins of herbal medicine using gas chromatography-mass spectrometry (GC-MS).

E0557: **Semi-parametric estimation of single-index models in modal regression**
Presenter: **Hirofumi Ohta**, The University of Tokyo, Japan
Semi-parametric estimation methods in modal regression, especially for single index models, are considered. The modal regression estimates the mode of the distribution of the outcome variable Y given regressors $X = x$ in the usual regression sense. Since conditional modes are defined by maximizers of conditional densities, non-parametric density estimation should be needed, so it occurs "the curse of dimensionality." To relax

this effect, we propose some semi-parametric estimation procedures for single index models, which are familiar with dimensionality reduction methods. Particularly in estimating weighted average derivatives of single index models, we use sample splitting technique which allows us to derive asymptotic theory under the machinery of U-statistics.

E0685: **A new clipping approach for parameter estimation in AR(p)**
Presenter: **Samuel Flimmel**, University of Economics in Prague, Czech Republic
Co-authors: Jiri Prochazka
With an increasing number of observations the probability of outlier presence also rises. This is a problem we face nowadays in many fields when working with big data. As it is known, standard methods are not able to work correctly with outliers, and, consequently, standard estimates are often biased. Therefore, sufficiently robust methods gain on importance. Autoregressive process AR(p) is well known and frequently used in statistics and economical modelling. One of the requirements for working with AR(p) models is an ability to estimate the parameters of the model correctly. In this poster, authors present a new robust method for parameter estimation in AR(p) models. The method is based on clipping the original time series and working with a binary time series instead. The clipping helps to deal with outliers, and, therefore, the estimation is not affected as much as when using a standard method. The new method is described and compared with existing robust methods using a simulation study performed in the R statistical software.

E0745: **Income inequality through ARDL modeling**
Presenter: **Alexandra Livada**, ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS, Greece
The aim is to investigate how and to what extent the main macroeconomic factors may affect income inequality as measured by top income shares. The macroeconomic indicators of most interest are the economic development and the openness of the economy, as well as education, financial development and inflation. The methodology of Autoregressive Distributed Lag (ARDL) cointegration procedure is employed to analyse empirically the long-run relationships and dynamic interaction among the variables of interest. The findings refer to the period before and after the economic crisis for northern European countries.

E0757: **On the prediction of the cross-section of returns in the long run from realized skewness and kurtosis**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic
Co-authors: Jozef Barunik
Large stream of literature has been dealing with factor-investing lately, however many of the supposedly significant asset pricing factors display poor out-of-sample performance. Therefore, we believe that it is important to incline to theory-based factors such as moments of returns distribution. Most popular moment-based factor is volatility, but there is a theoretical intuition and also empirical evidence, that investors are largely interested also in what happens in tails of distribution, i.e. in skewness and kurtosis. Moreover, the traditional asset pricing framework is working with over-the-frequencies aggregated information. If volatility, skewness and kurtosis are the right factors predicting the cross-section of returns, we can still get additional valuable insight by including information about different investment horizons of investors. We achieve this by using multiresolution analysis to construct short-term, medium-term and long-term measures of realized volatility, skewness and kurtosis. Then we construct an asset pricing model comprising these factors which is estimated using Fama-Macbeth type of regression.

E0803: **Variable selection on the mixture of additive quantile regressions model**
Presenter: **Wei-Te Lin**, National Dong Hwa University, Taiwan
Co-authors: Wei-Ying Wu
When observations come from the mixture of additive quantile regressions model, some unreasonable results of variable selection could happen if the existing quantile approaches are applied directly. We attempt to develop an algorithm to cluster data, select relevant variables, and identify the related structures simultaneously. In the developed algorithm, B-spline function is utilized to approximate the additive model and the quantile regression with the adaptively weighted group Lasso penalty is employed for the variable selection and structure detection. The performance of the proposed algorithm is discussed through simulation problems.

E0804: **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
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.

E0805: **Analysis of daily rainfall series in NW Spain: Searching for evidence of climatic change**
Presenter: **Maria Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Roland Fried, Jorge Marquinez
The aim is to analyze time series representing daily rainfall events collected in the Esva river basin in Asturias (northwest of Spain) during the last 50 years, where there is a lack of quantitative data about climatic oscillations. Our interest is a better understanding of the seasonal patterns found in these data and of possible trends which might indicate effects of climate change. An exploratory data analysis leads us to a first order Markov model with alternation between wet and dry periods. The probability of rainfall is modeled by logistic regression, and precipitation values on wet days are described by Gamma distributions with a first order autoregressive dependence structure. Lowess smoothing is used to include seasonality in both submodels because the seasonal pattern could not be covered adequately by standard periodic functions. A tentative result of our analysis is some evidence that dry periods get longer in the course of time, while precipitation values on wet days might be unaffected.

E0806: **Penalized multiple inflated values selection method with application to SAFER data**
Presenter: **Qiuya Li**, City University of Hong Kong, Hong Kong
Co-authors: Geoffrey Tso, Yichen Qin, Travis Lovejoy, Timothy Heckman, Yang Li
Expanding on the zero-inflated Poisson (ZIP) model, the multiple-inflated Poisson (MIP) model is applied to analyze count data with multiple inflated values. The existing studies on the MIP model determined the inflated values by inspecting the histogram of count response and fitting the model with different combinations of inflated values, which leads to relatively complicated computations and may overlook some real inflated points. We address a two-stage inflated values selection method, which takes all values of count response as potential inflated values and adopts the adaptive lasso regularization on the mixing proportion of those values. Numerical studies demonstrate the excellent performance both on inflated

values selection and parameters estimation. Moreover, a specially designed simulation, based on the structure of data from a randomized clinical trial of an HIV sexual risk education intervention, performs well and ensures our method could be generalized to the real situation. The empirical analysis of a clinical trial dataset is used to elucidate the MIP model.

E0811: **Detection of cell-type-specific risk-CpG sites in epigenome-wide association studies**

Presenter: **Xiangyu Luo**, The Chinese University of Hong Kong, Hong Kong

Co-authors: Can Yang, Yingying Wei

In epigenome-wide association studies, the measured signals for each sample are a mixture of methylation profiles from different cell types. The current approaches to the association detection only claim whether a cytosine-phosphate-guanine (CpG) site is associated with the phenotype or not, but they cannot determine the cell type in which the risk-CpG site is affected by the phenotype. Here, we propose a solid statistical method, HIgh REsolution (HIRE), which not only substantially improves the power of association detection at the aggregated level as compared to the existing methods but also enables the detection of risk-CpG sites for individual cell types.

Wednesday 20.06.2018 14:00 - 15:40 Parallel Session H – EcoSta2018

EO010 Room B4302 MODELLING FINANCIAL AND INSURANCE RISKS Chair: Tak Kuen Siu

E0702: **Stochastic Stackelberg differential reinsurance games**
Presenter: **Yang Shen**, York University, Canada
A new continuous-time framework is proposed to analyze optimal reinsurance, in which an insurer and a reinsurer are two players of a stochastic Stackelberg differential game, i.e., a stochastic leader-follower differential game. This allows us to determine optimal reinsurance from joint interests of the insurer and the reinsurer, which is rarely considered in a continuous-time setting. In the Stackelberg game, the reinsurer moves first and the insurer moves subsequently to achieve a Stackelberg equilibrium towards optimal reinsurance arrangement. Speaking more precisely, the reinsurer is the leader of the game and decides on optimal reinsurance premium to charge, while the insurer is the follower of the game and chooses optimal proportional reinsurance to purchase. We solve the game problem in two cases: exponential utility maximization and mean-variance optimization. Under the utility maximization framework, we find that the reinsurer always applies the variance premium principle to calculate the optimal reinsurance premium and the insurer's optimal ceding/retained proportion of insurance risk depends not only on the risk aversion of itself but also on that of the reinsurer. Under the mean-variance framework, if the reinsurer adopts the variance premium principle (resp. expected value premium principle), the Stackelberg equilibrium is attained when the insurer purchases the proportional reinsurance (resp. excess-of-loss reinsurance).

E0735: **A higher-order interactive hidden Markov model and its applications**
Presenter: **Wai-Ki Ching**, The University of Hong Kong, Hong Kong
A higher-order Interactive Hidden Markov Model(IHMM) is proposed, which incorporates both the feedback effects of observable states on hidden states and their mutual long-term dependence. The key idea of this model is to assume the probability laws governing both the observable and hidden states can be written as a pair of high-order stochastic difference equations. We also present an efficient procedure, a heuristic algorithm, to estimate the hidden states of the chain and the model parameters. Real applications in SSE Composite Index data and default data are given to demonstrate the effectiveness of our proposed model and corresponding estimation method.

E0742: **A higher-order Markov chain-modulated model for electricity spot-price dynamics**
Presenter: **Rogemar Mamon**, University of Western Ontario, Canada
Over the last three decades, the electricity sector worldwide underwent massive deregulation. As electricity is a non-storable commodity, its price is extremely sensitive to changes in supply and demand. The evolution of electricity prices exhibits pronounced mean reversion and cyclical patterns, possesses extreme volatility and relatively frequently occurring spikes, and manifests presence of state memory. We tackle the modelling and estimation problems under a new paradigm that integrates the deterministic calendar seasons and stochastic factors governing electricity prices. The de-seasonalised component of our proposed model has both the jump and mean-reverting properties to account for spikes and periodic cycles alternating between lower price returns and compensating periods of higher price returns. The parameters of the de-seasonalised model components are also modulated by a higher-order hidden Markov chain (HOHMC) in discrete time. The HOHMC's state is interpreted as the "state of the world" resulting from the interaction of various forces impacting the electricity market. Filters are developed to generate optimal estimates of HOHMC-relevant quantities using the observation process, and these provide online estimates of model parameters. We provide empirical demonstrations using the daily electricity spot prices compiled by the Alberta Electric System Operator.

E0744: **Time-consistent mean-variance reinsurance-investment problems under unbounded random parameters: BSDE and Uniqueness**
Presenter: **Hoi Ying Wong**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Bingyan Han
The open-loop time-consistent mean-variance (TCMV) reinsurance-investment problem is investigated when the parameters of the stochastic differential equations are stochastic and unbounded. The risk premium process of risky assets can be random and unbounded. Under an exponential integrability condition on the risk premium, we characterize the TCMV reinsurance-investment problem via a BSDE framework. An explicit solution to the equilibrium strategies is derived for a constant risk aversion under several popular financial models, including the constant elasticity of variance (CEV) and Ornstein-Uhlenbeck processes. For state-dependent risk aversion, a semi-closed form solution (up to the solutions to a pair of nonlinear PDEs) is obtained. Numerical results show that, under the CEV model, when stock price goes up, the equilibrium strategies suggest to invest more on the stock and less on the reinsurance protection. Under certain conditions, we prove the uniqueness of equilibrium strategies for constant and state-dependent risk aversion and this supplements results in the literature of using the HJB approach for the feedback controls.

EO105 Room G4701 NEW ALGORITHMS IN COMPLEX DATA ANALYSIS Chair: Samuel Mueller

E0181: **Sparse approximate inference for spatio-temporal point process models**
Presenter: **Andrew Zammit Mangion**, University of Wollongong, Australia
Co-authors: Botond Cseke, Guido Sanguinetti, Tom Heskes
Spatio-temporal log-Gaussian Cox process models play a central role in the analysis of spatially distributed systems in several disciplines. Yet, scalable inference remains computationally challenging both due to the high resolution modelling generally required and the analytically intractable likelihood function. We will demonstrate a novel way for solving this problem, which involves combining ideas from variational Bayes, message passing on factor graphs, expectation propagation, and sparse-matrix optimisation. The proposed algorithm is seen to scale well with the state dimension and the length of the temporal horizon with moderate loss in distributional accuracy. It hence provides a flexible and faster alternative to both non-linear filtering-smoothing type algorithms and approaches that implement the Laplace method (such as INLA) on (block) sparse latent Gaussian models. We demonstrate its implementation on simulation studies point-process observations, and use it to describe micro-dynamics in armed conflict in Afghanistan using data from the WikiLeaks Afghan War Diary.

E0324: **Multi-class modelling for muscle level prediction of beef eating quality**
Presenter: **Garth Tarr**, University of Sydney, Australia
Co-authors: Ines Wilms
The Meat Standards Australia (MSA) beef grading system was developed to improve the eating quality consistency of Australian beef for a range of muscles and cooking methods. The MSA system aims to predict the eating quality of beef given various inputs. Data has been collected and merged from experiments conducted over the past 20 years, resulting in a highly unbalanced data set where the majority of observations have been recorded on only a small subset of muscles with a far fewer observations on the majority of other muscles. The current approach to building a MSA grading system is ad hoc and labour intensive. We present a new method of multi-class modelling using lasso-type penalties to encourage similar coefficient estimates that bridge a spectrum of muscle level models. At one extreme, each muscle is modelled independently and at the other extreme a single pooled regression model for all muscles. An ideal model is somewhere between the two extremes, where muscles with limited

data are encouraged to borrow information from other similar muscles. We illustrate our method on real data and through simulation studies. The choice of tuning parameter is also discussed.

E0363: **Zero-inflated exponential families, with applications to count times-series**
Presenter: **Alan Huang**, University of Queensland, Australia
Co-authors: Thomas Fung
A novel mechanism is introduced that generates exponential families of zero-inflated distributions from a given nominal distribution. This proves particularly useful for observation-driven times-series modelling of zero-inflated counts, where neither the underlying zero-inflation nor count processes are fully observable. The idea is demonstrated on two data analysis examples.

E0612: **Fitting models with complex qualitative constraints**
Presenter: **Berwin Turlach**, The University of Western Australia, Australia
Several models are discussed for which the parameter space is constrained such that the boundary of the feasible space cannot be easily parameterised, but it is relatively easy to determine whether or not a set of parameters are feasible. This motivates the development of algorithms that can be used to fit such models to data.

EO218 Room LT-11 SEEMINGLY UNRELATED PAPERS IN NONPARAMETRIC ECONOMETRICS Chair: Daniel Henderson

E0318: **Nonparametric panel data models with cross-sectional dependence**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo, Peter Robinson
The asymptotic distribution for the local linear estimator in nonparametric panel data regression models is established when cross-sectional dependence is allowed. In order to take into account the information of the error covariance for estimates, a two step local linear regression technique is proposed. Sufficient conditions for its asymptotic normality are given and its efficiency gains relative to the standard nonparametric techniques is established. Asymptotically optimal bandwidth choices are justified for both estimates. Feasible optimal bandwidths, and feasible optimal regression estimates, are also asymptotically justified. The proposed estimators are augmented by a Monte Carlo study and they are also illustrated in an empirical analysis about the relationship between public debt, monetary policy and economic growth for eurozone countries.

E0177: **Nonparametric quantile regression for double censored data with application to stock markets with price limits**
Presenter: **Chi-Yang Chu**, National Taipei University, Taiwan
Quantile regression is often used in the analysis of stock return-volume relations. Many countries impose upper and lower limits on stock returns and losses, respectively, in order to reduce price volatility. However, this double censored property appears to be ignored in the literature. To analyze stock markets with price limits in a proper fashion, a nonparametric quantile regression model for double censored data is proposed. The proposed estimator performs well in simulations. In the application to Taiwanese stock markets, the proposed approach seems to alleviate some potential biases arising from double censored data. Specifically, the proposed estimator suggests larger estimated losses via conditional value at risk.

E0420: **Debiased machine learning for instrumental variable quantile regressions**
Presenter: **Jau-er Chen**, National Taiwan University, Taiwan
The aim is to investigate estimation and inference on a low-dimensional (causal) parameter in the presence of high-dimensional controls in an instrumental variable quantile regression. The estimation and inference are based on the Neyman-type orthogonal moment conditions, that are relatively insensitive to the estimation of the nuisance parameters. The Monte Carlo experiments show that the econometric procedure performs well. We also apply the procedure to empirically investigate the effect of 401(k) eligibility and participation on net financial assets.

E0410: **Exporting behavior and labor share in Chinese manufacturing industries: A semiparametric approach**
Presenter: **Jinjing Tian**, West Virginia University, United States
The purpose is to examine the relationship between exports and labor share in Chinese manufacturing industries, the most tradable sector in the economy. While most studies focus on the effect of trading with developing countries on domestic labor markets and/or labor share in developed economies, few studies examine its impact on labor in developing countries, such as China, the world's largest export economy. By using firm-level data in Chinese manufacturing industries and allowing for functional coefficients to take into account the heterogeneity of exporting behavior, we find that exports significantly decline labor's share of income neutrally through itself and non-neutrally through the channels of firms' capital intensity, technological progress, and monopoly power.

EO247 Room LT-12 NONPARAMETRIC APPROACHES FOR FUNCTIONAL AND HIGH-DIMENSIONAL DATA Chair: Zhengwu Zhang

E0510: **Reduced rank modeling for functional regression with functional responses**
Presenter: **Heng Lian**, City university of Hong kong, Hong Kong
Rgression problems are considered where both the predictor and the response are functional in nature. Driven by the desire to build a parsimonious model, we consider functional reduced rank regression in the framework of reproducing kernel Hilbert spaces, which can be formulated in the form of linear factor regression with estimated multivariate factors, and achieves dimension reduction in both the predictor and the response spaces. The convergence rate of the estimator is derived. Simulations and real data sets are used to demonstrate the competitive performance of the proposed method.

E0523: **A nonparametric Bayesian model for clustering inhomogeneous Poisson processes**
Presenter: **Xiaowei Wu**, Virginia Tech, United States
Co-authors: Hongxiao Zhu
Random events arise in many applications and numerous data have been generated to record their distribution in time or space. A commonly used model for such data is the inhomogeneous Poisson process (IHPP), which is characterized by its time- or location-dependent intensity function. Motivated by a genomic application of identifying transcriptional regulatory modules using modern ChIP-seq data, we developed a nonparametric Bayesian clustering model for samples of multiple IHPPs. This model, called Dirichlet process mixture of log Gaussian Cox process (DPM-LGCP), employs a DP prior to the random distribution of the latent IHPP log intensity functions to facilitate clustering of functional data and the consequent IHPPs arising therefrom. To overcome the inference difficulty caused by calculating marginal likelihood of IHPP, we adopt approximate Bayesian inference based on integrated nested Laplace approximations (INLA), and integrate the approximated marginal likelihood into DPM sampling. Simulation studies show that DPM-LGCP achieves good accuracy and robustness and outperforms two alternative clustering methods. We apply this model to learn transcription factor binding patterns and identify transcriptional regulatory modules in mouse ES cells. Findings from such analysis help uncover how transcription factors work together in orchestrating gene activity in ES cells, and will eventually lead to a better understanding of different tissue development and disease progression.

E0499: **Nonparametric Bayes models of fiber curves connecting brain regions**
Presenter: **Zhengwu Zhang**, University of Rochester, United States
In studying structural inter-connections of human brain, it is common to first estimate fiber bundles connecting different regions relying on diffusion MRI. These fiber bundles act as highways for neural activity, snaking through the brain and connecting different regions. Current statistical methods for analyzing these fibers reduce the rich information into an adjacency matrix, with the elements containing a count of fibers or a mean diffusion feature along the fibers. The goal is to avoid discarding the rich geometric information of fibers, developing flexible models for characterizing the population distribution of fibers between brain regions of interest within and across different individuals. We start by decomposing each fiber into a corresponding rotation matrix, shape and translation from a global reference curve. These components can be viewed as data lying on a product space composed of different Euclidean spaces and manifolds. To non-parametrically model the distribution within and across individuals, we rely on a hierarchical mixture of product kernels specific to the component spaces. Taking a Bayesian approach to inference, we develop efficient methods for posterior sampling. The approach automatically produces clusters of fibers within and across individuals, and yields interesting new insights into variation in fiber curves.

E0793: **Sparse Poisson regression with penalized weighted score function**
Presenter: **Lihu Xu**, University of Macau, China
A new penalized method is proposed in order to solve sparse Poisson Regression problems. Being different from l1 penalized log-likelihood estimation, our new method can be viewed as penalized weighted score function method, which possesses a tuning-free feature. We show that under mild conditions, our estimator is l1 consistent and the tuning parameter can be pre-specified, which enjoys the same good property as the square-root Lasso. Besides, we conduct numerical simulations and apply our method into image reconstruction. All the results exhibit our method has a better performance than others.

EO057 Room LT-14 NEW DEVELOPMENT OF FUNCTIONAL DATA ANALYSIS Chair: Lily Wang

E0322: **Spatially varying coefficient model for functional image-on-scalar regression**
Presenter: **Lily Wang**, Iowa State University, United States
Co-authors: Guannan Wang
Motivated by recent work analyzing imaging data in the neuroimaging studies, a class of linear functional response regression models for imaging responses and scalar predictors is considered. To help define the possibly irregular domain of the active (non-null) region, our smoothing method is based on penalized bivariate splines over triangulations, which solves the problem of "leakage" across non-rectangular domains where many conventional smoothing tools suffer. Highly efficient and scalable estimation algorithm is developed. We conduct Monte Carlo simulation to examine the finite-sample performance of the proposed method. An Alzheimer's Disease Study example will be provided to illustrate the application of the proposed methods.

E0349: **Multi-threshold accelerated failure time model**
Presenter: **Jialiang Li**, NUS, Duke-NUS, SERI, Singapore
A two-stage procedure for simultaneously detecting multiple thresholds and achieving model selection in the segmented accelerated failure time (AFT) model is developed. In the first stage, we formulate the threshold problem as a group model selection problem so that a concave 2-norm group selection method can be applied. In the second stage, the thresholds are finalized via a refining method. We establish the strong consistency of the threshold estimates and regression coefficient estimates under some mild technical conditions. The proposed procedure performs satisfactorily in our simulation studies. Its real world applicability is demonstrated via analyzing a follicular lymphoma data.

E0347: **Two-sample functional linear models**
Presenter: **Hua Liang**, George Washington University, United States
Two-sample functional linear regression with a scaling transformation of regression functions are studied. We consider estimation of the intercept, the slope function and the scalar parameter based on the functional principal component analysis. We also establish the rates of convergence for the estimator of the slope function, which is shown to be optimal in a minimax sense under certain smoothness assumptions. We further investigate semiparametric efficiency for the estimation of the scalar parameter and hypothesis testing. We also extend the proposed method to sparsely and irregularly sampled functional data and establish the consistency for the estimators of the scalar and the slope function. We evaluate numerical performance of the proposed methods through simulation studies and illustrate their utility via analysis of an AIDS data set.

E0470: **Binary classification of functional data via continuously additive modeling**
Presenter: **Yichao Wu**, The University of Illinois at Chicago, United States
Co-authors: Kai Shen, Hans-Georg Mueller, Fang Yao
The continuously additive model was recently proposed as a new nonlinear functional regression technique. It lends great flexibility to the study of functional data. To explore the use of continuously additive modeling for functional classification, we propose to couple it with the support vector machine, a large margin classifier, aiming at the classification of functional data. The support vector machine is a popular binary classification method that has enjoyed great success but has not become popular yet for the important task of classifying functional data. The support vector machine has been shown to be sensitive to outliers since it is based on an unbounded hinge loss. Therefore we couple the continuously additive modeling technique with the robust support vector machine using the truncated hinge loss as well. We illustrate the performance of our methods with simulation examples and two data sets that involve the classification of spectral data. The proposed approach is compared with classification based on the functional linear model.

EO111 Room LT-15 METHODS FOR MODELING SPATIO-TEMPORAL DATA Chair: Won Chang

E0417: **A Bayesian spatial market segmentation method using Dirichlet process-Gaussian mixture models**
Presenter: **Won Chang**, University of Cincinnati, United States
Co-authors: Sunghoon Kim, Heewon Chae
A new spatial market segmentation method is proposed by using a Bayesian spatial Dirichlet process-Gaussian mixture (SDPGM) model. The approach forms segments of spatial locations that show a similar relationship between service factors and customer satisfaction. Our method employs a SDPGM model in the spatial domain and hence automatically determines the number of segments and the membership of spatial locations simultaneously in a unified framework. We also incorporate ridge and lasso regularization in parameter estimation to better select statistically significant service factors, which is important for efficient resource allocation in marketing. Our simulation study confirms that (i) the proposed approach can successfully identify the hidden spatial segmentation structure only based on the spatial distribution and the predictor-response relationship and (ii) the ridge estimator shows a better performance in identifying non-significant variables and hence leads to a better identification of key market drivers than lasso. We apply the proposed approach to an online customer satisfaction data set for the restaurants in Washington DC area collected by Yelp. The results provide an interesting insight on the key drivers of customer satisfaction in different sub-regions of the area.

E0387: **Bayesian functional ANOVA models in climate prediction**
Presenter: **Xia Wang**, University of Cincinnati, United States
The aim is to construct a Bayesian functional ANOVA model, which is efficient in computation and flexible in structure for complex, temporally and spatially correlated, large-scale climate data. Current ANOVA models for spatiotemporal data limits the capability of data ensembles from multiple sources. The proposed model combines data from different climate models in a structured way, providing not only an ensemble of the variable of interest for climate prediction but also a spatially and temporally varying term for RCM and GCM model comparison, which is not available in current research and is of great interests to climatologists and modelers. Specifically, the model assumes that the observational data and the simulation results from different computer models share the same unobserved underlying process, while the observational data has noises and the simulation outputs from climate models contains a model-dependent discrepancy, which is then decomposed into three components: the discrepancy caused by RCM difference, the discrepancy caused by GCM different, and the interaction effects of RCM and GCM. Modeling of these components is particularly challenging when extending to the high-dimensional, complex dependence structure. The proposed model increases the accuracy in climate prediction as well as improving the scientists understanding and assessment of different climate models.

E0431: **A unified exposure prediction approach for multivariate spatial data**
Presenter: **Roman Jandarov**, University of Cincinnati College of Medicine, United States
Co-authors: Zheng Zhu
When presented with a multi-pollutant exposure problem, it is possible to adapt univariate methods by applying them independently to each pollutant. While these advanced statistical models for predicting pollution exposures can incorporate all important meteorological, geographical and land-use information and allow for spatiotemporal dependence between the pollution concentrations at close distances in space and time to obtain accurate exposure predictions for a single pollutant, applying the univariate models to each pollutant of the multi-pollutant scenario will ignore a potentially important source of information that lies in the correlations between the pollutants. We develop novel unified exposure prediction approaches for multi-pollutant data based on the idea of linking models in a chain. In the proposed approaches, we apply univariate models sequentially and the predicted exposures from each model are used in the subsequent models as an input. We also incorporate dimension reduction and variable selection techniques before applying each model. The methods are applied to simulated data with different covariance structures and to monitoring data from U.S. Environmental Protection Agency. The results demonstrate that chain-based approaches can lead to increased prediction accuracy compared to traditional univariate models.

E0695: **Hierarchical Bayesian autoregressive models in South Korea ozone**
Presenter: **Sanghoo Yoon**, Daegu University, Korea, South
Co-authors: Dain Park
Environmental has a property of space-time. Having both the spatial and temporal dimensions adds substantial complexity to environmental data analysis. We model daily maximum 8-hour ozone concentration data obtained from $n = 32$ sites in South Korea for analysis between 2013 and 2017. Maximum temperature, relative humidity, and wind speed were considered as covariates. The data on the square root scale seems most attractive in terms of both symmetry and stabilizing the variance. Independent Gaussian process model and autoregressive model are specified within a hierarchical Bayesian framework and Markov Chain Monte Carlo techniques. These space-time models allow accurate spatial prediction of a temporally aggregated ozone summary along with its uncertainty.

EO194 Room LT-16 RECENT ADVANCES IN BAYESIAN METHODS Chair: Antonio Lijoi

E0178: **Bayesian variable selection under misspecified errors**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
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.

E0408: **A Bayesian nonparametric spiked process prior for dynamic model selection**
Presenter: **Michele Guindani**, University of California, Irvine, United States
In many applications, investigators consider processes that vary in space and time, with the goal of identifying temporally persistent and spatially localized departures of those processes from a baseline or "normal" behavior. We propose a Bayesian nonparametric model selection approach for the analysis of spatio-temporal data, which takes into account the non-exchangeable nature of measurements collected over time and space. More specifically, a zero-inflated conditionally identically distributed (CID) species sampling prior is used to model temporal dependence in the selection, by borrowing information across time and assigning data to clusters associated to either a null or an alternate process. Spatial dependences are accounted for by means of a Markov random field (MRF) prior, which allows to inform the selection based on inferences conducted at nearby locations. We investigate the performances of our model by means of a simulation study and an application to a disease surveillance problem, for detecting outbreaks of pneumonia and influenza (P&I) mortality in the continental United States. We show how the proposed modeling framework compares favorably with respect to commonly adopted threshold methods for detecting outbreaks over time and also to recent proposals modeling more complex Markov switching dependences.

E0193: **Scalable Bayesian nonparametric clustering and classification**
Presenter: **Yang Ni**, UT Austin, United States
A scalable multi-step Monte Carlo algorithm is developed for inference under (possibly non-conjugate) Bayesian nonparametric models. Each step is "embarrassingly parallel" and can be implemented using the same Markov chain Monte Carlo sampler. The simplicity and generality of our approach makes a wide range of Bayesian nonparametric methods applicable to large datasets. Specifically, we apply product partition model with regression on covariates using novel implementation to classify and cluster patients in large electronic health records study. We find interesting clusters and superior classification performance against competing classifiers.

E0428: **Estimating clusters from multivariate binary data via hierarchical Bayesian Boolean matrix factorization**
Presenter: **Zhenke Wu**, University of Michigan, United States
Co-authors: Livia Casiola-Rosen, Antony Rosen, Scott Zeger
An ongoing challenge in subsetting autoimmune disease patients is to define autoantibody signatures produced against a library of elemental molecular machines each comprised of multiple component autoantigens. It is of significant value to quantify both components of the machines

and the striking variations in their frequencies among individuals. Based on multivariate binary responses that represent subject-level presence or absence of proteins over a grid of molecular weights, we develop a Bayesian hierarchical model that represents observations as aggregation of a few unobserved machines where the aggregation varies by subjects. Our approach is to specify the model likelihood via factorization into two latent binary matrices: machine profiles and individual factors. Given latent factorization, we account for inherent uncertainties in immunoprecipitation, errors in measurement or both using sensitivities and specificities of protein detection. The posterior distribution for the numbers of patient clusters and machines are estimated from data and by design tend to concentrate on smaller values. The posterior distributions of model parameters are estimated via Markov chain Monte Carlo which makes a list of molecular machine profiles with uncertainty quantification as well as patient-specific posterior probability of having each machine. We demonstrate the proposed method by analyzing patients gel electrophoresis autoradiography (GEA) data for patient subsetting.

EO061 Room LT-17 RECENT DEVELOPMENTS IN TIME SERIES ANALYSIS AND INSURANCE Chair: Sangyeol Lee

E0187: **Asymptotic properties of mildly explosive processes with locally stationary disturbance**
Presenter: **Junichi Hirukawa**, Niigata University, Japan
Co-authors: Sangyeol Lee
For a class of locally stationary process, we consider the problem of testing ARMA model against other non-stationary ARMA model. When testing the problem, we use linear serial rank statistics and contiguity of LeCam's notion. If the null hypothesis is white noise, then, under the null and the alternative, the asymptotic normality of the proposed statistics is established by using the locally asymptotic normality (LAN). We incorporate the locally stationary phenomena in the testing problem.

E0219: **Prediction intervals for time series and their applications to portfolio selection**
Presenter: **Shih-Feng Huang**, National University of Kaohsiung, Taiwan
Co-authors: Hsiang-Ling Hsu
The aim is to consider prediction intervals for time series and to apply the results to portfolio selection. The dynamics of the high and low underlying returns are depicted by time series models, which lead to a prediction interval of future returns. We propose an innovative criterion for portfolio selection based on the prediction interval. A new concept of coherent risk measures for the interval of returns is introduced. An empirical study is conducted with the stocks of the Dow Jones Industrial Average Index. A self-financing trading strategy is established by daily reallocating the holding positions via the proposed portfolio selection criterion. The numerical results indicate that the proposed prediction interval has promising coverage, efficiency, and accuracy for prediction. The proposed portfolio selection criterion constructed from the prediction intervals is capable of suggesting an optimal portfolio according to the economic conditions.

E0378: **Modeling the dependence in compound model using copula representation when the size of frequency is informative**
Presenter: **Jae Youn Ahn**, Ewha Womans University, Korea, South
Co-authors: Woojoo Lee, Rosy Oh
While the copula method is a popular choice in modelling the dependence, the choice of the proper copula family is much harder in general compared to the choice of the proper marginal distribution families. Especially, in the modelling of dependence between frequency and average severity, we show, by example, that classical copula approach may mislead the dependence between frequency and severity. We provide the copula model which can safely model the dependence between frequency and individual severity and the dependence between severities. Since the proposed model can distinguish the dependence between frequency and individual severity and the dependence between severities, the model can be applied to other field where the cluster size, the frequency in actuarial context, is informative.

E0463: **Valuing equity-indexed annuities with icicled barrier options**
Presenter: **Bangwon Ko**, Soongsil University, Korea, South
Inspired by the recent popularity of autocallable structured products, the purpose is to enhance equity-indexed annuities (EIAs) by introducing a new class of barrier options, termed icicled barrier options. The new class of options has a vertical (icicled) barrier along with the horizontal one of the ordinary barrier options, which may act as an additional knock-in or knock-out trigger. To improve the crediting method of EIAs, we propose a new EIA design, termed autocallable EIA, with payoff structure similar to the autocallable products except for the minimum guarantee, and further investigate the possibility of embedding various icicled barrier options into the plain point-to-point or the ratchet EIAs. Explicit pricing formulas for the proposed EIAs and the icicled barrier options are obtained under the Black-Scholes model. To the purpose, we derive the joint distribution of the Brownian motion at the icicled time and the maturity, and its running maximum. As an application of the well-known reflection principle, the derivation itself is an interesting probability problem and the joint distribution plays a key role in the subsequent pricing stage. Our option pricing result can be easily transferred to EIAs or other equity-linked products. The pricing formulas for the EIAs and the options are illustrated through numerical examples.

EO320 Room LT-18 ADVANCES IN STATISTICAL MODELLING FOR COMPLEX BIOMEDICAL AND HEALTH DATA Chair: Shu-Kay Ng

E0465: **Semiparametric transformation models for interval-censored data in the presence of a cure fraction**
Presenter: **Chyong-Mei Chen**, National Yang-Ming University, Institute of Public Health, Taiwan
Mixed case interval censored data arises when the event of interest is known only to occur within an interval induced by a sequence of random examination times. Such data are commonly encountered in disease research with longitudinal follow-up. Furthermore, the medical treatment has progressed over the last decade with an increasing proportion of patients being cured for many types of diseases. Thus, interest has grown in cure models for survival data which hypothesizes a certain proportion of subjects in the population are not expected to experience the events of interest. A two-component mixture cure model for regression analysis of mixed interval censored data is considered. The first component is a logistic regression model that describes the cure rate, and the second component is a semiparametric transformation model that describes the distribution of event time for the uncured subjects. Semiparametric maximum likelihood estimation for the considered model is proposed. An EM type algorithm is developed for obtaining the semiparametric maximum likelihood estimators (SPMLE) of regression parameters and establishes the large sample properties. Extensive simulation studies indicate that the SPMLE performs satisfactorily in a wide variety of settings. A medical study is provided for illustration.

E0494: **Joint model for bivariate zero inflated recurrent event data with terminal event**
Presenter: **Yang-Jin Kim**, Sookmyung Women University, Korea, South
Multivariate recurrent events arise when a subject has experienced several types of event repeatedly over time. In many situations, a substantial portion of subjects have no events even during a fairly long follow-up time. To reflect such event free group, zero inflated model has been implemented in a content of recurrent event. Similar situation is incorporated as a cure rate model in an ordinary survival analysis and the most commonly used method is a mixture model where two possibilities are simultaneously considered. The motivation comes from Korean long-term care insurance data, where two types of recurrent events occur. The former is a hospitalization and the latter is an out-patient services. Subject can experience two events repeatedly until either death or censoring. However, about 60 percent of participants had not experience these events. A

joint model is suggested to model regression models for two recurrent events with zero inflation as well as terminal event. Some simulation studies are performed to evaluate the biases of the suggested models.

E0479: **Mixture modelling of high-dimensional data**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia
Some aspects of the use of finite mixture distributions in modelling high-dimensional data are considered. Attention is focussed on mixtures with multivariate normal and t-distributions and some skew variants for the component distributions. Dimension reduction is undertaken via factor models which allow for skew distributions in addition to white noise for the factor distributions. Consideration is also given to dimension reduction via clustering of the variables. Applications are given involving the analyses via mixture modelling of some real data sets in the biomedical and health sciences.

E0539: **Sensitivity analysis for publication bias in meta-analysis of diagnostic studies for a continuous biomarker**
Presenter: **Satoshi Hattori**, Osaka University, Japan
Publication bias is one of the most important issues in meta-analysis. For standard meta-analyses to examine intervention effects, the funnel plot and the trim-and-fill method are simple and widely used techniques for assessing and adjusting for the influence of publication bias, respectively. However, their use may be subjective and can then produce misleading insights. To make a more objective inference for publication bias, various sensitivity analysis methods have been proposed, including the Copas selection model. For meta-analysis of diagnostic studies evaluating a continuous biomarker, the summary receiver operating characteristic (sROC) curve is a very useful method in the presence of heterogeneous cutoff values. To our best knowledge, no methods are available for the evaluation of influence of publication bias on the estimation of the sROC curve. We introduce a Copas-type selection model for meta-analysis of diagnostic studies and propose a sensitivity analysis method for publication bias. The proposed method enables us to assess the influence of publication bias on the estimation of the sROC curve and then judge whether the result of the meta-analysis is sufficiently confident or should be interpreted with much caution. We illustrate our proposed method with real data.

EO259 Room P4302 BAYESIAN METHODS IN NETWORK ANALYSIS Chair: Peter Orbanz

E0648: **Subsampling and inference for beta neutral-to-the-left models of random graphs**
Presenter: **Benjamin Bloem-Reddy**, University of Oxford, United Kingdom
Beta Neutral-to-the-Left (NTL) models are able to generate random graphs with any level of sparsity, and with power law degree distributions of any possible exponent. This flexibility is unique among models in the recent literature based on various notions of exchangeability, which are able to obtain sparsity values and power law exponents over a limited range of possible values. The flexibility of Beta NTL models comes at the cost of losing any obvious form of exchangeability. However, by conditioning on (or inferring) the vertex arrival times, a constrained notion of exchangeability becomes apparent, and may be exploited for efficient estimation and inference algorithms, as well as for model-coherent subsampling, even when no ordering information is available.

E0661: **Bayesian extensions of neural network-based graphon approximations**
Presenter: **Creighton Heaukulani**, Goldman Sachs, Hong Kong
The perspective of directly modeling the graphon of an exchangeable graph has, naturally, been approached recently with neural network function approximations. Desirable properties in many statistical network applications, such as block structure (i.e., community detection) and dynamic extensions are straightforwardly accomplished using tools from deep learning. We will discuss these two extensions, in particular, and see that inference for our suggested models demands a Bayesian approach. Several applications in the context of finance will be demonstrated.

E0617: **Function estimation on a large graph using Bayesian Laplacian regularization**
Presenter: **Alisa Kirichenko**, CWI (Centrum Wiskunde & Informatica), Netherlands
Co-authors: Harry Zanten
In recent years there has been substantial interest in high-dimensional estimation and prediction problems on large graphs. These can in many cases be viewed as high-dimensional or nonparametric regression or classification problems in which the goal is to learn a "smooth" function on a given graph. We present a mathematical framework that allows us to study the performance of nonparametric function estimation methods on large graphs and we derive minimax convergence rates within the framework. We consider simple undirected graphs that satisfy an assumption on their "asymptotic geometry", formulated in terms of the graph Laplacian. We also introduce a Sobolev-type smoothness condition on the target function using the graph Laplacian to quantify smoothness. Then we develop Bayesian procedures for problems at hand and we show how asymptotically optimal Bayesian regularization can be achieved under these conditions. The priors we study are randomly scaled Gaussians with precision operators involving the Laplacian of the graph.

E0663: **Nonparametric models for structured sparse graphs**
Presenter: **Sinead Williamson**, University of Texas at Austin, United Kingdom
There has been recent interest in the Bayesian community in models for sparse graphs with an unbounded number of vertices. Such models are appropriate for modeling large social or interaction networks, where the number of vertices scales approximately linearly with the number of interactions. However, sparsity is only one aspect of the structure of such networks, and naive sparse models tend to ignore the presence of locally dense sub-graphs and latent communities. We propose models appropriate for binary and integer-valued graphs that are globally sparse, but which contain locally dense sub-graphs, and show how these models can be used to infer latent communities from social network data.

EO014 Room P4701 NEW ADVANCES IN STATISTICAL COMPUTING AND COMPLEX DATA ANALYSIS Chair: Tsung-I Lin

E0337: **Measuring financial interdependence in asset returns with an application to Euro Zone equities**
Presenter: **Cody Yu-Ling Hsiao**, Macau University of Science and Technology, China
Co-authors: Renee Fry-McKibbin, Vance Martin
A general procedure is proposed to identify changes in asset return interdependence over time using entropy theory. The approach provides a decomposition of interdependence in terms of co-moments including co-skewness, co-kurtosis and co-volatility as well as more traditional measures based on second order moments such as correlations. A new diagnostic test of independence is also developed which incorporates these higher order co-moments. The properties of the entropy interdependence measure are demonstrated using a number of simulation experiments, as well as applying the methodology to euro zone equity markets over the period 1990 to 2017.

E0501: **Correcting for differential recruitment with respondent-driven sampling data**
Presenter: **Isabelle Beaudry**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Krista Gile
Respondent-driven sampling (RDS) is a sampling mechanism that has proven effective to sample hard-to-reach human populations connected through social networks, such as certain populations at higher risk of HIV/AIDS infection. Under RDS, a small number of individuals known to the researcher are initially sampled and asked to recruit a fixed small number of their contacts who belong to the target population. Subsequent sampling

waves are produced by peer recruitment until the desired sample size is achieved. However, the researcher's lack of control over the sampling process has posed a number of challenges to valid statistical inference. For instance, it is often assumed that participants recruit completely at random among their contacts. However, participants may systematically over recruit individuals with a particular characteristic or with whom they engage in a given type of relationship. Literature suggests that most RDS prevalence estimators are greatly sensitive to this assumption. In this work, we define three forms of differential recruitment, provide methods for estimating them and introduce novel inferential methods to address the induced biases.

E0599: **Bayesian semiparametric modeling for HIV longitudinal data with censoring and skewness**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Victor Hugo Lachos Davila, Wan-Lun Wang, Vanda Inacio
In biomedical studies, the analysis of longitudinal data based on Gaussian assumptions is common practice. Nevertheless, more often than not, the observed responses are naturally skewed, rendering the use of symmetric mixed effects models inadequate. In addition, it is also common in clinical assays that the patients responses are subject to some upper and/or lower quantification limit, depending on the diagnostic assays used for their detection. Furthermore, the responses may also often present a nonlinear relation with some covariates such as time. To address the aforementioned three situations, we consider a Bayesian semiparametric model based on a combination of splines and wavelets for longitudinal censored data using the multivariate skew-normal distribution. The proposed semiparametric approach is focused on the use of splines to approximate the nonlinear general mean and wavelets for modeling the individual trajectories per subject. Lastly, the use of the skew normal distribution allows to capture the skewness of the data. The newly developed method is illustrated through simulated data and real data concerning AIDS/HIV infected patients.

E0531: **Mixtures of common restricted skew-t factor analyzers**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan
Co-authors: Wan-Lun Wang, Luis Mauricio Castro Cepero
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. The aim is to present a further robust extension of the MCFA and MCtFA models, called the mixture of common restricted skew-t factor analyzers (MCrstFA), by assuming a restricted multivariate skew-t distribution for the common factors. The MCrstFA model can be used to accommodate severely non-normal random phenomena while preserving its parsimony in factor-analytic representation and performing graphical visualization in low-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 our proposed model is illustrated with simulated and real datasets, and experimental results signify its superiority over some existing competitors.

EO200 Room P4703 RECENT ADVANCES IN INCOMPLETE DATA ANALYSIS Chair: Alex Kin Yau Wong

E0230: **Outcome-dependent sampling with interval-censored failure time data**
Presenter: **Qingning Zhou**, University of North Carolina at Charlotte, United States
Epidemiologic studies and disease prevention trials often seek to relate an exposure variable to a failure time that suffers from interval-censoring. When the failure rate is low and the time intervals are wide, a large cohort is often required so as to yield reliable precision on the exposure-failure-time relationship. However, large cohort studies with simple random sampling could be prohibitive for investigators with a limited budget, especially when the exposure variables are expensive to obtain. Alternative cost-effective sampling designs and inference procedures are therefore desirable. We propose an outcome-dependent sampling (ODS) design with interval-censored failure time data, where we enrich the observed sample by selectively including certain more informative subjects. We develop semiparametric likelihood approaches for analyzing data from the proposed interval-censoring ODS design. The consistency and asymptotic normality of the resulting regression parameter estimators are established. The results from our extensive simulation study show that the proposed design and method works well for practical situations and is more efficient than the alternative designs and competing approaches. An example from the Atherosclerosis Risk in Communities (ARIC) study is provided for illustration.

E0244: **Semiparametric optimal estimation with nonignorable nonresponse data**
Presenter: **Kosuke Morikawa**, Osaka University and The University of Tokyo, Japan
Co-authors: Jae Kwang Kim
When the response mechanism is believed to be not missing at random (NMAR), a valid analysis requires stronger assumptions on the response mechanism than standard statistical methods would otherwise require. Semiparametric estimators have been developed under the model assumptions on the response mechanism. A new identification condition is proposed to guarantee model identifiability without using any instrumental variable. Furthermore, we develop optimal semiparametric estimation for parameters such as the population mean. Specifically, we propose two semiparametric optimal estimators that do not require any model assumptions other than the response mechanism. Asymptotic properties of the proposed estimators are discussed. An extensive simulation study is presented to compare with some existing methods. We present an application of our method using Korean Labor and Income Panel Survey data.

E0355: **Generalization of Heckman selection model to nonignorable nonresponse using call-back information**
Presenter: **Baojiang Chen**, University of Texas Health Science Center at Houston – Austin Regional Campus, United States
Call-back of nonrespondents is common in surveys involving telephone or mail interviews. In general, these call-backs gather information on unobserved responses, so incorporating them can improve the estimation accuracy and efficiency. Call-back studies mainly focus on Alho's selection model or the pattern mixture model formulation. We generalize the Heckman selection model to nonignorable nonresponses using call-back information. The unknown parameters are then estimated by the maximum likelihood method. The proposed formulation is simpler than Alho's selection model or the pattern mixture model formulation. It can reduce the bias caused by the nonignorably missing mechanism and improve the estimation efficiency by incorporating the call-back information. Further, it provides a marginal interpretation of a covariate effect. Moreover, the regression coefficient of interest is robust to the misspecification of the distribution. Simulation studies are conducted to evaluate the performance of the proposed method. For illustration, we apply the approach to National Health Interview Survey data.

E0580: **Partition-based screening with ultrahigh-dimensional data**
Presenter: **Yi Li**, University of Michigan, United States
Traditional variable selection methods are compromised by overlooking useful information on covariates with similar functionality or spatial proximity, and by treating each covariate independently. Leveraging prior grouping information on covariates, we propose partition-based screening methods for ultrahigh-dimensional variables in the framework of generalized linear models. We show that partition-based screening exhibits the sure screening property with a vanishing false selection rate, and we propose a data-driven partition screening framework with unavailable or unreliable prior knowledge on covariate grouping and investigate its theoretical properties. We consider two special cases: correlation-guided partitioning and spatial location-guided partitioning. In the absence of a single partition, we propose a theoretically justified strategy for combining statistics from

various partitioning methods. The utility of the proposed methods is demonstrated via simulation and analysis of functional neuroimaging data.

EO028 Room P4704 SURVIVAL AND COUNT DATA ANALYSIS Chair: Geoffrey Tso

E0464: **JobEnomics: Firm growth prediction by online job posting data**
Presenter: **Jing Wu**, City University of Hong Kong, Hong Kong
What would you do if you had access to over 44m unique job postings representing approximately 28,000 distinct private and public companies with over 32,000 new jobs posted daily as well as 16bn words captured in job posting descriptions? In this novel research, we explain firm growth and human capital investment using a giant web-crawled textual dataset on job posting. Our results show that job posting information has strong predictive power on firm future performance and stands out differently from investment patterns such as CMA in Fama French five-factor asset pricing model.

E0306: **Integrative gene-gene interaction analysis for high dimensional data**
Presenter: **Yang Li**, Renmin University of China, China
For many complex diseases, extensive omics profiling has been extensively conducted. It has been shown that gene-gene interactions may have important implications beyond the main genetic effects. The number of unknown parameters in a gene-gene interaction analysis is usually much larger than the sample size. As such, results generated from analyzing a single dataset are often unsatisfactory. Integrative analysis, which jointly analyzes the raw data from multiple independent studies, has been conducted in a series of recent studies and shown to outperform single-dataset analysis, meta-analysis, and other multi-datasets analyses. The goal is to conduct integrative analysis in the identification of gene-gene interactions. For regularized estimation and selection of important interactions (and main effects), we apply a Threshold Gradient Directed Regularization (TGDR) approach. Advancing from the exiting studies, the TGDR approach is modified to respect the "main effects, interactions" hierarchy. The proposed approach has an intuitive formulation and is computationally simple and broadly applicable. Simulations and the analysis of cancer prognosis data with gene expression measurements demonstrate its satisfactory practical performance.

E0259: **Multiscale risk forecasting: A deep learning based ensemble approach**
Presenter: **Kaijian He**, Hunan University of Science and Technology, China
Co-authors: Geoffrey Tso, Yingchao Zou
In the volatile financial markets characterized by complex mixture of different underlying dynamics, a new multiscale ensemble approach has been proposed for forecasting more accurate value at risk measures. It takes advantage of a new multiscale technique called variational mode decomposition as the basis for disensembling the underlying risk factors in the multiscale domain. The individual characteristics of these risk factors are modeled using different econometrics and artificial intelligence models. The forecasts for these risk factors serve as the ensemble members. Based on that, the nonlinear ensemble model is employed to aggregate the ensemble members and produce the optimal forecasts. The time varying dynamic weights for the ensemble members are modeled using the deep learning model, including the widely popular Long Short Term Memory (LSTM) etc. Empirical evaluation of the performance of the proposed model have been conducted using the extensive database, constructed by daily price observations in the major crude oil markets, including West Taxes Intermediate and Brent crude oil markets. Experiment results confirm that the proposed forecasting models produce an improved forecasting accuracy for the risk estimates. It is found that different risk factors has different time scale focus. Their influence on the joint risk movement is time varying and dynamic, which linear ensemble models fail to capture.

E0232: **A new survival model based on extended diffusion theory**
Presenter: **Lianlian Song**, Nanjing University of Aeronautics and Astronautics, China
Co-authors: Geoffrey Tso
The mixed pressure is considered from three aspects and a new survival model is developed based on the diffusion theory. However, different from the conventional diffusion theory that considers the adoption patterns of innovators and imitators, we extend it by dividing the imitators into two separate groups as original imitators and competition imitators. The original imitators are those who purchase the observed brand according to the social information of observed brand users, and the competition imitators are those who purchase the observed brand, but influenced by the social information of competitive brand users. In the actual market, multiple competitive brands exist. The internal pressure of the social system is not only from the previous adopters of the observed brand, but also from the adopters of competitive brands. Examining the competition imitators may indicate the competitive power of the observed brand, and supply more information for consumer behavior study and marketing strategy. Finally, we compare and confirm the model fit with other two benchmark models using the data from an E-commerce brand company.

Wednesday 20.06.2018 16:10 - 17:50 Parallel Session I – EcoSta2018

EO055 Room B4302 RECENT ADVANCES IN MODELLING FINANCE DATA AND RISK ASSESSMENT Chair: Charles Au

E0369: **Statistical properties of the modified multivariate skew-t distribution**
Presenter: **Charles Au**, University of Sydney, Australia
Co-authors: Boris Choy
A class of distributions known as the multivariate skew-normal distribution has been a popular choice for capturing skewness and correlation in multivariate data. A heavy-tailed version of this distribution, known as the skew-t distribution, is often used for data with fat tails, such as asset returns. A more general extension to the skew-t distribution, the modified multivariate skew-t distribution (Mod-skew-t distribution), is proposed. It is more flexible in that the degrees of freedom parameter of each of the marginal distributions can be allowed to vary. This overcomes the limitation that these parameters must be the same, as is the case for the non-modified version of the skew-t distribution. Using the fact that the Mod-skew-t distribution has the scale mixtures of skew-normal (SMSN) representation, Bayesian inference will be used for parameter estimation. Its various statistical properties and application to statistical modelling will be explored.

E0395: **A semi-parametric realized joint quantile regression framework for financial tail risk forecasting**
Presenter: **Chao Wang**, The University of Sydney, Australia
Co-authors: Richard Gerlach
A new realized joint Value at Risk (VaR) and expected shortfall (ES) quantile regression framework is proposed, through incorporating a measurement equation into the joint quantile regression model. The measurement equation models the contemporaneous dependence between the realized measures (e.g. Realized Variance and Realized Range) and the latent conditional quantile. Further, sub-sampling and scaling methods are applied to both the realized range and realized variance, to help deal with inherent micro-structure noise and inefficiency. An adaptive Bayesian Markov Chain Monte Carlo method is employed for estimation and forecasting, whose properties are assessed and compared with maximum likelihood through simulation study. In a forecasting study applied to 7 market indices and 2 individual assets, compared to a range of parametric, non-parametric and semi-parametric models, including GARCH, Realized-GARCH, CARE and the original joint VaR and ES quantile regression models, one-day-ahead Value-at-Risk and Expected Shortfall forecasting results favor the proposed models, especially when incorporating the sub-sampled Realized Variance and the sub-sampled Realized Range.

E0492: **Stock/bond volatility/correlation on macro factors in China: Based on GARCH-MIDAS**
Presenter: **Qian Chen**, Peking University Shenzhen Campus, China
Co-authors: Chen Chen
The GARCH-MIDAS-X models are applied to China stock and bond market in attempt to examine the power of low-frequency macro factors in predicting high-frequency market volatility. The results confirm the significant relationship between the macro variables and the long run volatility. Specifically, the long-run component of GARCH-MIDAS model incorporating industrial added value growth rate (IP) accounts for around 30% of total conditional volatility of Chinas stock market and bond market. The study also finds that, industrial added value is a better predictor than the producer price index, which may be due to the fact that the China economy is still in the development stage and the market is more sensitive to economic growth than inflation. The out-of-sample forecasts of GARCH-MIDAS-X models improve with longer horizons. Though for the stock market, GARCH-MIDAS-RV still performs best in semi-annual horizon; for bond, GARCH-MIDAS with IP volatility outperforms other models in semi-annual horizon. DCC-MIDAS-X is also applied to study the relationship between the macro factors and the stockbond correlation. The results suggest a weaker effect of the macro factors, which may be due to the absence of inter-market macro-strategy investors in China.

E0791: **Advanced statistical models for cryptocurrency research**
Presenter: **Jennifer Chan**, University of Sydney, Australia
Cryptocurrencies as of late have commanded global attention on a number of fronts: heightened pecuniary, technological infrastructures, and investment interest. Cryptocurrency is different from fiat currency in many ways: no institutional control, near instantaneous transaction and a fast changing cryptocurrency community. Yet there are many controversies surrounding cryptocurrency, its monetary role, price formulation, investment devices, etc. As cryptocurrency is increasingly accepted even by some major banks, we aim to derive advanced time series models to understand the statistical properties of cryptocurrency, including its notoriously wild volatility. This speculative nature of cryptocurrency has raised debates over its monetary role against speculative asset. We highlight some stylised facts about the volatility dynamic including its oscillatory persistence and leverage effect and relate these results to their respective cryptographic designs. The data analysis is initiated with a broad scope of cryptocurrencies, then a more detailed understanding of the top 5 by market capitalization and followed up with a specific focus on Bitcoin. The results favour Gegenbauer long memory over standard long memory filter to model the logarithm of the daily range.

EO115 Room G4302 DEVELOPMENTS IN MACROECONOMIC FORECASTING Chair: Tatsuma Wada

E0312: **Normality tests for latent variables**
Presenter: **Dante Amengual**, CEMFI, Spain
Co-authors: Enrique Sentana, Tincho Almuzara
The rationale behind the Louis formula is exploited to derive simple to implement and interpret score tests of normality in the innovations to the latent variables in linear state space models against generalized hyperbolic alternatives, including symmetric and asymmetric Student t's. We decompose our tests into third and fourth moment components, and obtain one-sided likelihood ratio analogues, whose asymptotic distribution we provide. When we apply them to a cointegrated dynamic factor model which combines the expenditure and income versions of US aggregate real output to improve its measurement, we reject normality if the sample period extends beyond the Great Moderation.

E0335: **A transfer entropy test for causality in longitudinal data**
Presenter: **Andres Romeu**, Universidad de Murcia, Spain
Co-authors: Maximo Camacho, Manuel Ruiz-Marin
Granger causality (GC) is the most popular statistical definition of causal relationship. In this type of analysis avoiding sample selection problems often requires data from many individuals and many time periods, whose observations are pooled in cross section or collected in panel data. However, this testing procedure could give a misleading account of the causality effect due to typical data problems. In particular, the size and power of this test can be seriously affected when the linearity assumption breaks down, such as under random coefficients, heterogeneous panels, structural breaks, and extreme observations. We propose a causality test based on the concept of transfer entropy between two variables that is simultaneously robust against these data problems. The test uses the pooled sample to disclose causality between the variables of interest of a generic form without having to impose an ex-ante structure of heterogeneity in the causal relationship. A Monte Carlo experiment with five different scenarios shows that the test displays correct size and high power to detect causality in situations where linear GC fails. We provide two empirical examples. One uses per capita GDP and government expenditure yearly World Bank data in a panel of 100 countries for the 1961-2016

period. Second, we test causality between firm size and productivity using data from the Bureau of Labor Statistics on 86 manufacturing firms for the 1988-2015 period.

E0506: **Multivariate Bayesian predictive synthesis in macroeconomic forecasting**
Presenter: **Kenichiro McAlinn**, University of Chicago, United States
Co-authors: Knut Are Aastveit, Jouchi Nakajima, Mike West
The methodology and a detailed case study in use of a class of Bayesian predictive synthesis (BPS) models for multivariate time series forecasting is developed. This extends the recently introduced foundational framework of BPS to the multivariate setting, with detailed application in the topical and challenging context of multi-step macroeconomic forecasting in a monetary policy setting. BPS evaluates sequentially and adaptively over time varying forecast biases and facets of miscalibration of individual forecast densities, and critically of time-varying inter-dependencies among them over multiple series. We develop new BPS methodology for a specific subclass of the dynamic multivariate latent factor models implied by BPS theory. Structured dynamic latent factor BPS is here motivated by the application context sequential forecasting of multiple US macroeconomic time series with forecasts generated from several traditional econometric time series models. The case study highlights the potential of BPS to improve of forecasts of multiple series at multiple forecast horizons, and its use in learning dynamic relationships among forecasting models or agents.

E0438: **An alternative estimation method for time-varying parameter models**
Presenter: **Tatsuma Wada**, Keio University, Japan
Co-authors: Mikio Ito, Akihiko Noda
A non-Bayesian, regression-based or generalized least squares (GLS)- based approach is formally proposed to estimate a class of time-varying AR parameter models. This approach has partly been used previously, and is proven to be efficient because, unlike conventional methods, it does not require Kalman filtering and smoothing procedures, but yields a smoothed estimate that is identical to the Kalman-smoothed estimate. Unlike the maximum likelihood estimator, the possibility of the pile-up problem is negligible. In addition, this approach enables us to deal with stochastic volatility models, models with a time-dependent variance-covariance matrix, and models with non-Gaussian errors that allow us to deal with abrupt changes or structural breaks in time-varying parameters.

EO279 Room G4701 SOME MODERN TOPICS RELATED TO SPATIAL STATISTICS Chair: Yumou Qiu

E0561: **Replicated spatial temporal data models**
Presenter: **Yuzhen Zhou**, University of Nebraska Lincoln, United States
Co-authors: Honglang Wang, Yingjie Li
Statistical models are considered for spatial temporal data with independent realizations, such as the functional magnetic resonance imaging (fMRI) data. A semiparametric covariance structure is proposed to model the space-time dependence. Functional principle component methods together with non-linear least squares methods are applied to model estimations. The inference of mean functions under the proposed space-time dependence structure is investigated. We offer simulation experiments to evaluate our proposed models.

E0633: **High dimensional discriminant analysis for spatially dependent data**
Presenter: **Yingjie Li**, Michigan State University, United States
Co-authors: Taps Maiti
Linear discriminant analysis(LDA) is one of the most classical and popular classification techniques. However, it performs poorly in high-dimensional classification. Many sparse discriminant methods have been proposed to make LDA applicable in high dimensional case. One drawback of those methods is the structure of the covariance among features is ignored. We propose a new procedure for high dimensional discriminant analysis for spatially correlated data. Penalized maximum likelihood estimation (PMLE) is developed for feature selection and parameter estimation. Tapering technique is applied to reduce computation load. The theory shows that the method proposed can achieve consistent parameter estimation, features selection, and asymptotically optimal misclassification rate. Extensive simulation study and real data analysis shows a significant improvement in classification performance under spatial dependence.

E0626: **Mean test for China haze pollution data over time**
Presenter: **Shuyi Zhang**, Peking University, China
Co-authors: Yumou Qiu, Songxi Chen
Motivated by the evaluation of China haze pollution over time, we consider a two-sample mean test for high-dimensional time series data. A U-statistic is proposed to estimate the sum of squared difference between the population means. Its limiting distribution is derived under the null and alternative hypotheses. The proposed method is free of distribution assumption, and is adaptive to temporal dependence. It also does not require specific relationship between sample size and dimension. Simulation study is carried out to compare the performance of the proposed test with other existing methods. Applications to testing significance of China PM2.5 annual difference is discussed.

E0688: **Inference on multi-level brain connectivities based on fMRI data**
Presenter: **Yumou Qiu**, University of Nebraska Lincoln, United States
The brain functional network models coherent activities between different brain regions, which plays an important role in our cognition and behavior. We propose a hierarchical model on partial correlations to study the brain functional interactivity based on multi-subject functional magnetic resonance imaging (fMRI) data. Multiple testing procedures adaptive to time dependent data with false discovery proportion control are proposed to identify both the population and subject specific brain connectivities. A computationally feasible algorithm is developed. Theoretical results and simulation studies demonstrate the good properties of the proposed procedures. A real example on fMRI data from normal healthy persons and patients with Parkinson's disease shows that several brain connections are missing in Parkinson patients.

EO089 Room LT-11 BIG DATA IN FINANCE Chair: Ningning Xia

E0427: **Solving the Markowitz optimization problem for large portfolios**
Presenter: **Mengmeng Ao**, Xiamen University, China
Co-authors: Yingying Li, Xinghua Zheng
The large dimensional Markowitz optimization problem is studied. Given any risk constraint level, we introduce a new approach for estimating the optimal portfolio. The approach relies on a novel unconstrained regression representation of the mean-variance optimization problem, combined with high-dimensional sparse regression methods. Our estimated portfolio, under a mild sparsity assumption, asymptotically achieves mean-variance efficiency and meanwhile effectively controls the risk. To the best of our knowledge, this is the first approach that can achieve these two goals simultaneously for large portfolios. The superior properties of our approach are demonstrated via comprehensive simulation and empirical studies.

E0512: **Dynamic change-detection with application to mutual fund selection**
Presenter: **Lilun Du**, HKUST, China
Co-authors: Changliang Zou
In the era of big data, it has become particularly important to rapidly and sequentially identify individuals whose behavior deviates from the norm. In such applications, the state of a stream can alternate, possibly multiple times, between a null and an alternative state. Aiming to balance the ability to detect two types of changes, i.e., a change from the null to the alternative and back to the null, we develop a large-scale dynamic testing system in the framework of false discovery rate (FDR) control. By fully exploiting the sequential feature of datastreams, we propose a new procedure based on a penalized version of the generalized likelihood ratio test statistics for change-detection. The FDR at each time point is shown to be controlled under some mild conditions on the dependence structure of datastreams. A data-driven approach for choosing the penalization parameter is developed, giving the new method an edge over existing methods in terms of FDR control and detection delay. Its advantage is demonstrated using a real data example in Chinese financial market.

E0645: **The bet for similarity: Adaptive discrete smoothing with application in finance**
Presenter: **Ye Luo**, University of Florida, United States
Co-authors: Martin Spindler, Victor Chernozhukov, Xi Chen
The traditional linear panel data model does not allow individuals to have different predictive models except for different intercepts. This restriction is too strong to analyze modern data with heterogeneity in individual behaviors. We consider a scenario where individuals are allowed to have their own models. In practice, such a setting leads to undesirable prediction because the longitude of the data is usually short. We consider a set of algorithm called Adaptive Discrete Smoothing (or the ADS algorithm). The ADS algorithm tries to cluster similar individuals and utilize the data from the clustered group for prediction. This method is beneficial when the cross-sectional size of the panel data is large. The ADS algorithm can be applied to non-linear panel data, high-dimensional panel data and etc. We show that the ADS algorithm can substantially improve the theoretical convergence rate as well as the practical prediction performance in panel data. We apply the ADS algorithm to the Fama-French regression framework with monthly return data in the U.S. stock market.

E0440: **Testing high-dimensional covariance matrices under the elliptical distribution and beyond**
Presenter: **Xinxin Yang**, ISOM, HKUST, Hong Kong
Co-authors: Xinghua Zheng, Jiaqi Chen, Hua Li
High-dimensional covariance matrices testing is studied under a generalized elliptical model. The model accommodates several stylized facts of real data including heteroskedasticity, heavy-tailedness, asymmetry, etc. We consider the high-dimensional setting where the dimension p and the sample size n grow to infinity proportionally, and establish a central limit theorem (CLT) for the linear spectral statistic (LSS) of the sample covariance matrix based on self-normalized observations. The CLT is different from the existing ones for the LSS of the usual sample covariance matrix. Our tests based on the new CLT neither assume a specific parametric distribution nor involve the kurtosis of data. Simulation studies show that our tests work well even when the fourth moment does not exist. Empirically, we analyze the idiosyncratic returns under the Fama-French three-factor model for SP 500 Financials sector stocks, and our tests reject the hypothesis that the idiosyncratic returns are uncorrelated.

EO214 Room LT-12 RANDOM PROJECTION APPROACHES TO HIGH-DIMENSIONAL STATISTICAL PROBLEMS Chair: Timothy Cannings

E0191: **On the use of random projections for dimension reduction in linear regression**
Presenter: **Martin Slawski**, George Mason Univ, United States
Principal Components Regression (PCR) is a traditional tool for dimension reduction in linear regression that has been both criticized and defended. One concern about PCR is that obtaining the leading principal components tends to be computationally demanding for large data sets. While Random Projections (RPs) do not possess the optimality properties of the projection onto the leading principal subspace, they are computationally appealing and hence have become increasingly popular in recent years. We present an analysis showing that the dimension reduction offered by RPs achieves a prediction error in subsequent regression close to that of PCR, at the expense of requiring a slightly large number of RPs than PCs.

E0584: **Classification in the presence of label noise: Structure-aware error bounds via random projections**
Presenter: **Henry Reeve**, University of Birmingham, United Kingdom
Co-authors: Ata Kaban
New error bounds for linear classification in conditions of random label noise are presented. To obtain the bounds we use random projections as an analytic device to gain advantage from benign geometric structure that may be present in the data, while working directly with the 0-1 loss. The bounds are data dependent, highlight the characteristics of the problem that make the problem easier or harder, and remain informative in small sample conditions.

E0651: **Classification algorithm based on random iterated projections**
Presenter: **Zhengdao Wang**, Iowa State University, United States
Co-authors: Qi Xiao
Random projections are useful in reducing the dimensionality of a data set while almost preserving the distances between the data points. For this reason, they have found uses in many dimensionality-reduction applications. Algorithms have been proposed that use random projections for classification problems. In such algorithms, the projections are usually applied to the data points in one batch to map the data points to a lower dimensional space. We will present an algorithm that performs that projection in a sequential manner. The subsequent projection vectors, although dependent on the vectors used in earlier projections, are not required to be orthogonal to them. We will present the rationale for using such sequential projection, and also present numerical results that compare the proposed algorithm with existing methods.

E0694: **Supervised random projection T test for two sample test in high dimension**
Presenter: **Radhendushka Srivastava**, Indian Institute of Technology Bombay, India
In high dimension, testing the equality of mean of two populations have been explored by several researchers. The RAndom Projection T Test (RAPTT) is an exact test for equality of means of two normal populations. The RAPTT does not require constraints on the dimension and the sample size. However, the random projection inserts additional randomness in the decision making. We use supervised random projection based on the data and propose Supervised RAndom Projection T test (SRAPTT). We illustrate the advantage of supervised through simulation and gene expression data.

EO324 Room LT-13 NONLINEAR FINANCIAL ECONOMETRICS Chair: Jeroen Rombouts

E0709: **Index and individual stock term structure of variance risk premia**
Presenter: **Jeroen Rombouts**, ESSEC Business School, France
For the individuals stocks 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.

E0713: **Dynamic properties and correlation structure of a large panel of cryptocurrencies**
Presenter: **Francesco Violante**, ENSAE ParisTech, France
Co-authors: Jeroen Rombouts, Luc Bauwens
The behaviour of a large portfolio of highly valued and most actively traded cryptocurrencies is studied. Unlike more traditional financial assets, the dynamic behaviour of cryptocurrencies returns is characterised by a particularly high level of volatility, by abnormally large variations, and is affected by extreme shocks to liquidity. We aim at investigating the dynamic properties of cryptocurrencies and particularly the correlation structure linking them to identify whether and to what extent there exist diversification opportunities in these markets.

E0741: **Geographic dependence and diversification in house price returns: The role of leverage**
Presenter: **Andreas Heinen**, Universite de Cergy Pontoise, France
Co-authors: Mi Lim Kim
The aim is to analyze time variation in the average dependence within a set of regional monthly house price index returns in a regime switching multivariate copula model with a high and a low dependence regime. Using equidependent Gaussian copulas, we show that the dependence of house price returns varies across time, which reduces the gains from the geographic diversification of real estate and mortgage portfolios. More specifically, we show that a decrease in leverage, and to a lesser extent an increase in mortgage rates, is associated with a higher probability of moving to and staying in the high dependence regime.

E0770: **Pricing individual stock options using both stock and market index information: New results**
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 factor. 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. To address these shortcomings, we model the evolution of the individual stock returns together with the market index returns in a flexible bivariate model that allows us to estimate risk premia in line with the theory. We assess the performance of the model by pricing individual stock options on the constituent stocks in the Dow Jones Industrial Average over a long time period including the recent Global Financial Crisis.

EO083 Room LT-14 DATA, MODELS, LEARNING AND BEYOND Chair: Catherine Liu

E0708: **Subgroup selection in adaptive signature designs of confirmatory clinical trials**
Presenter: **Zhiwei Zhang**, University of California at Riverside, United States
The increasing awareness of treatment effect heterogeneity has motivated flexible designs of confirmatory clinical trials that prospectively allow investigators to test for treatment efficacy for a subpopulation of patients in addition to the entire population. If a target subpopulation is not well characterized in the design stage, it can be developed at the end of a broad eligibility trial under an adaptive signature design. New procedures for subgroup selection and treatment effect estimation (for the selected subgroup) are proposed under an adaptive signature design. We first provide a simple and general characterization of the optimal subgroup that maximizes the power for demonstrating treatment efficacy or the expected gain based on a specified utility function. This characterization motivates a procedure for subgroup selection that involves prediction modelling, augmented inverse probability weighting and low dimensional maximization. A cross-validation procedure can be used to remove or reduce any resubstitution bias that may result from subgroup selection, and a bootstrap procedure can be used to make inference about the treatment effect in the subgroup selected. The approach proposed is evaluated in simulation studies and illustrated with real examples.

E0724: **Semiparametric Bayesian analysis for longitudinal mixed effects models with non-normal AR(1) errors**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Junshan Shen, Jin Yang
The focus is on Bayesian inference on the longitudinal mixed effects model with non-normal AR(1) errors. We model the nonparametric zero-mean noise in the autoregression residual by the Dirichlet process (DP) mixture model. Applying the empirical likelihood tool, an adjusted sampler based on the Polya urn representation of DP is proposed to incorporate information of the moment constraints of the mixing distribution. Gibbs sampling algorithm based on the adjusted sampler is proposed to approximate the posterior distributions under DP priors. The proposed method can be easily extended to deal with other moment constraints owing to the wide application background of empirical likelihood. Simulation studies evaluate the performance of the proposed method. Our method is illustrated in analysis of a longitudinal data set from a psychiatric study.

E0726: **Quantile regression for functional partially linear models in ultra-high dimensions**
Presenter: **Haiqiang Ma**, Jiangxi University of Finance and Economics, China
A functional partially linear quantile model in ultra-high dimensional scenarios is considered, where the response is scalar and the predictors include both multiple random processes and high-dimensional scalar covariates. A framework of regularization with two nonconvex penalty functions in the context of functional partially linear quantile regression is proposed formally, and the selection and estimation of important variables can be then achieved by minimizing a double penalized functional quantile objective function. In a theoretical investigation, we establish the asymptotic properties of the resulting estimators based on the difference convex analysis (DCA) under some regularity conditions, and also consider the convergence rate of the prediction of the conditional quantile function. The empirical performance and the usefulness of our proposed approach are demonstrated through a large number of simulation studies and a real application.

E0728: **Automatic shape-constrained nonparametric regression**
Presenter: **Huixia Wang**, The George Washington University, United States
Co-authors: Yanlin Tang, Zhikun Gao
The vocalizations of mice consist of syllables of different types determined by the frequency modulations and structure variations. To characterize the impact of social environments and genotypes on vocalizations, it is important to identify the shapes of frequency contours of syllables. Using hypothesis testing methods to determine the shapes would require testing various null and alternative hypotheses for each curve, and is impractical for vocalization studies with a large number of frequency contours. To overcome this challenge, we propose a new penalization-based method, which provides function estimation and automatic shape identification simultaneously. The method estimates the functional curve through quadratic B-spline approximation, and captures the shape feature by penalizing the positive and negative parts of the first two derivatives of the spline function in a group manner. Under some regularity conditions, we show that the proposed method can identify the correct shape with probability approaching one, and the resulting nonparametric estimator can achieve the optimal convergence rate. Simulation shows the proposed method gives more stable curve estimation than the unconstrained B-spline estimator, and it is competitive to the shape-constrained estimator assuming prior knowledge of the functional shape. The proposed method is applied to the motivating vocalization study to examine the effect of Mecp2 gene on the vocalizations of mice during courtship.

EO063 Room LT-15 CONTEMPORARY BAYESIAN INFERENCE FOR HIGH-DIMENSIONAL MODELS Chair: Richard Gerlach

E0433: **Inversion copulas for GARCH models and tail risk forecasting**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Michael Smith, Worapree Ole Maneesoonthorn
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 proposed 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.

E0435: **Variational Bayes estimation of time series copulas for multivariate ordinal and mixed data**
Presenter: **Ruben Loaiza Maya**, University of Melbourne, Australia
Co-authors: Michael Smith
A new variational Bayes method for estimating high-dimensional copulas with discrete, or discrete and continuous, margins is proposed. The method is based on a variational approximation to a tractable augmented posterior, and is substantially faster than previous likelihood-based approaches. We use it to estimate drawable vine copulas for univariate and multivariate Markov ordinal and mixed time series. These have dimension rT, where T is the number of observations and r is the number of series, and are difficult to estimate using previous methods. The vine pair-copulas are carefully selected to allow for heteroskedasticity, which is a common feature of ordinal time series data. When combined with flexible margins, the resulting time series models also allow for other common features of ordinal data, such as zero inflation, multiple modes and under- or over-dispersion. Using data on homicides in New South Wales, and also U.S bankruptcies, we illustrate both the flexibility of the time series copula models, and the efficacy of the variational Bayes estimator for copulas of up to 792 dimensions and 60 parameters. This far exceeds the size and complexity of copula models for discrete data that can be estimated using previous methods.

E0447: **Implicit copulas from Bayesian regularized regression smoothers**
Presenter: **Michael Smith**, University of Melbourne, Australia
Co-authors: Nadja Klein
The aim is to show how to extract the implicit copula of a response vector from a Bayesian regularized regression smoother with Gaussian disturbances. The copula can be used to compare smoothers that employ different shrinkage priors and function bases. We illustrate with three popular choices of shrinkage priors a pairwise prior, the horseshoe prior and a g prior augmented with a point mass as employed for Bayesian variable selection and both univariate and multivariate function bases. To evaluate the implicit copula we first construct a Gaussian copula by conditioning on the regularization parameters, and then mix over them using numerical or Monte Carlo methods. This greatly simplifies computation of the implicit copula compared to direct evaluation. The copulas are combined with non-parametric margins to extend the regularized smoothers to non-Gaussian data. Efficient Markov chain Monte Carlo schemes for evaluating the copula are given for this case. Using both simulated and real data, we show how such copula smoothing models can improve the quality of resulting function estimates and predictive distributions.

E0718: **High-dimensional ABC**
Presenter: **Yanan Fan**, University of New South Wales, Australia
Approximate Bayesian computation (ABC) has emerged in recent years as a popular alternative for Bayesian inference in the presence of an intractable likelihood. The idea is primarily based on sampling, and the use of summary statistics. We will discuss several ideas of doing ABC in high dimensions.

EO145 Room LT-17 PREDICTIVE ANALYTICS AND TIME SERIES ANALYSIS Chair: Cathy W-S Chen

E0307: **Realized stochastic volatility models with generalized Gegenbauer long memory**
Presenter: **Manabu Asai**, Soka University, Japan
In recent years fractionally differenced processes have received a great deal of attention due to their flexibility in financial applications with long memory. We develop a new realized stochastic volatility (RSV) model with general Gegenbauer long memory (GGLM), which encompasses a new RSV model with seasonal long memory (SLM). The RSV model uses the information from returns and realized volatility measures simultaneously. The long memory structure of both models can describe unbounded peaks apart from the origin in the power spectrum. For estimating the RSV-GGLM model, we suggest estimating the location parameters for the peaks of the power spectrum in the first step, and the remaining parameters based on the Whittle likelihood in the second step. We conduct Monte Carlo experiments for investigating the finite sample properties of the estimators, with a quasi-likelihood ratio test of RSV-SLM model against the RSV-GGLM model. We apply the RSV-GGLM and RSV-SLM model to three stock market indices. The estimation and forecasting results indicate the adequacy of considering general long memory.

E0226: **Unit root testing for the Buffered autoregressive model**
Presenter: **Wai-Keung Li**, The University of Hong Kong, Hong Kong
Co-authors: Di Wang
Buffered (hysteretic) auto-regression is an extension of the classical threshold autoregressive model by allowing a buffered region for regime changes. We study asymptotic statistical inference for the 2-regime buffered autoregressive (BAR) model with possible unit roots. A sup-LR test

is proposed for testing for the nonlinear buffer effect in the possible presence of unit roots, and a class of unit root tests is proposed to identify the number of nonstationary regimes under the BAR model. The wild bootstrap is suggested to approximate the critical values of the tests. Two real examples are considered to illustrate the proposed methodology.

E0398: **Random weighting the Portmanteau tests for multivariate white noise with unknown dependent structure**
Presenter: **Muyi Li**, Xiamen University, China
The Ljung-Box portmanteau test is one of the most popular model diagnostic tools. However, when the errors are not i.i.d.random variables, the classical portmanteau test does not follow asymptotically chi-squared distribution. We employ the random weighting method to bootstrap the critical values of Ljung-Box portmanteau test in multivariate time series models with unknown dependent white noise. A set of Monte Carlo experiments demonstrate the practical relevance of this method. Real examples on the USD-MYR and USD-SGD five-day exchange rates illustrates the merits of our testing procedures.

E0556: **Doubly constrained factor models with applications to multivariate time series analysis**
Presenter: **Henghsiu Tsai**, Academia Sinica, Taiwan
The focus is on factor analysis of multivariate time series. We propose statistical methods that enable analysts to leverage their prior knowledge or substantive information to sharpen the estimation of common factors. Specifically, we consider a doubly constrained factor model that enables analysts to specify both row and column constraints of the data matrix to improve the estimation of common factors. The row constraints may present classifications of individual subjects whereas the column constraints may show the categories of variables. We derive both the maximum likelihood and least squares estimates of the proposed doubly constrained factor model and use simulation to study the performance of the analysis in finite samples. Real data are used to demonstrate the application of the proposed model.

EO314 Room LT-18 COMPUTATION CHALLENGES IN STATISTICAL METHODS Chair: Teng Zhang

E0204: **Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow**
Presenter: **Kean Ming Tan**, University of Minnesota, United States
Co-authors: Zhaoran Wang, Han Liu, Tong Zhang
Sparse generalized eigenvalue problem plays a pivotal role in a large family of high-dimensional learning tasks, including sparse Fishers discriminant analysis, canonical correlation analysis, and sufficient dimension reduction. However, existing methods and theory in the context of specific statistical models require restrictive structural assumptions on the input matrices. We exploit a nonconvex optimization perspective to study the sparse generalized eigenvalue problem under a unified framework. In particular, we propose the truncated Rayleigh flow method (Rifle) to estimate the leading generalized eigenvector and show that it converges linearly to a solution with the optimal statistical rate of convergence. Theoretically, our method significantly improves upon the existing literature by eliminating the structural assumption on the input matrices. To achieve this, our analysis involves two key ingredients: (i) a new analysis of the gradient based method on nonconvex objective functions, as well as (ii) a fine-grained characterization of the evolution of sparsity patterns along the solution path. Thorough numerical studies are provided to back up our theory.

E0388: **Adaptive basis sampling for smoothing splines**
Presenter: **Nan Zhang**, Fudan University, China
Smoothing splines provide flexible nonparametric regression estimators. However, the high computational cost of smoothing splines for large datasets has hindered their wide application. We develop a new method, named adaptive basis sampling, for efficient computation of smoothing splines in super-large samples. Smoothing spline for a regression problem with sample size n can be expressed as a linear combination of n basis functions and its computational complexity is generally of cubic n order. We achieve a more scalable computation in the multivariate case by evaluating the smoothing spline using a smaller set of basis functions, obtained by an adaptive sampling scheme that uses values of the response variable. Our asymptotic analysis shows that smoothing splines computed via adaptive basis sampling converge to the true function at the same rate as full basis smoothing splines. We show that the proposed method outperforms a sampling method that does not use the values of response variables in several applications.

E0507: **Provable convex co-clustering of tensors**
Presenter: **Eric Chi**, North Carolina State University, United States
Co-authors: Brian Gaines, Will Wei Sun, Hua Zhou
Clustering is a fundamental unsupervised learning technique that aims to discover groups of objects in a dataset. Biclustering extends clustering to two dimensions where both observations and variables are grouped simultaneously, such as clustering both cancerous tumors and genes or both documents and words. We develop and study a convex formulation of the generalization of biclustering to co-clustering the modes of multiway arrays or tensors, the generalization of matrices. Our convex co-clustering (CoCo) estimator is guaranteed to obtain a unique global minimum of the formulation and generates an entire solution path of possible co-clusters governed by a single tuning parameter. We extensively study our method in several simulated settings, and also apply it to an online advertising dataset. We also provide a finite sample bound for the prediction error of our CoCo estimator.

E0578: **Constrained regression via majorization-minimization**
Presenter: **Jason Xu**, University of California Los Angeles, United States
The majorization-minimization (MM) principle generalizes expectation-maximization (EM) algorithms to settings beyond missing data. Like EM, the idea relies on transferring optimization of a difficult objective (i.e. the likelihood under missing data) to a sequence of simpler subproblems (i.e. maximizing the expectation of the likelihood under complete data). We discuss MM approaches to regression problems under constraints such as sparsity and low-rankness, and simple recipes for building the family of surrogate functions to be iteratively optimized. Through this lens, we revisit sparse covariance estimation and high-dimensional regression. We present strong empirical performance on several data examples and convergence guarantees even for non-convex objectives.

EO235 Room P4302 RECENT ADVANCES IN SOCIAL NETWORK ANALYSIS Chair: Rui Pan

E0171: **A note on estimating network dependence in a discrete choice model**
Presenter: **Jing Zhou**, Renmin University of China, China
Co-authors: Da Huang, Hansheng Wang

Discrete choice model is probably one of the most popularly used statistical methods in practice. The common feature of this model is that it considers the behavioral factors of a person and the assumption of independent individuals. However, this widely accepted assumption seems problematic, because human beings do not live in isolation. They interact with each other and form complex networks. Then, the application of discrete choice model to network data will allow for network dependence in a general framework. We focus on a discrete choice model with probit error which is specified as a latent spatial autoregressive model(SAR). This model could be viewed as a natural extension of the classical SAR model. The key difference is that the network dependence is latent and unobservable. Instead, it could be measured by a binary response variable. Parameter estimation then becomes a challenging task due to the complicated objective function. Following the idea of composite likelihood, an approximated paired maximum likelihood estimator (APMLE) is developed. Numerical studies are carried out to assess the finite sample performance of the proposed estimator. Finally a real dataset of Sina Weibo is analyzed for illustration purpose.

E0245: **A popularity scaled latent space model for large-scale directed social network**
Presenter: **Danyang Huang**, Renmin University of China, China
Co-authors: Xiangyu Chang, Hansheng Wang

Large-scale directed social network data often involve degree heterogeneity, reciprocity, and transitivity properties. A sensible network generating model should take these features into consideration. To this end, we propose a popularity scaled latent space model for the large-scale directed network structure formulation. It assumes for each node a position in a hypothetically assumed latent space. Then, the nodes close (far away) to each other should have larger (less) probability to be connected. As a consequence, the reciprocity and transitivity properties can be analytically derived. In addition to that, we assume for each node a popularity parameter. Those nodes with larger (smaller) popularity are more (less) likely to be followed by other nodes. By assuming different distributions for popularity parameters, different types of degree heterogeneity can be modeled. Furthermore, based on the proposed model, a comprehensive probabilistic index is constructed for link prediction. Its finite sample performance is demonstrated by extensive simulation studies and a Sina Weibo (a Twitter-type social network in China) dataset. The performances are competitive.

E0477: **Spatio-temporal autoregressions with network information**
Presenter: **Yingying Ma**, Beihang University, China
Co-authors: Shaojun Guo, Qiwei Yao

A new class of spatio-temporal models with unknown autoregressive coefficient matrices is proposed. The setting represents a sparse structure for high-dimensional spatial panel dynamic models when panel members represent economic (or other type) individuals at many different locations. Due to the innate endogeneity, we apply two different approaches to estimate the autoregressive coefficient matrices. A variable selection method has been developed for determining the unknown elements in the autoregressive matrices. Some asymptotic properties of the inference methods are established. The proposed methodology is further illustrated using both simulated and real data sets.

E0558: **Grouped network vector autoregression**
Presenter: **Rui Pan**, Central University of Finance and Economics, China

In time series analysis, it is of great interest to model a continuous response for all the individuals at equally spaced time points. With the rapid advance of social network sites, network data are becoming increasingly available. In order to incorporate the network information among individuals, a network vector autoregression (NAR) model has been recently developed. The response of each individual can be explained by its lagged value, the average of its neighbors, and a set of node-specific covariates. However, all the individuals are assumed to be homogeneous since they share the same autoregression coefficients. To express individual heterogeneity, we develop a grouped NAR (GNAR) model. Individuals in a network can be classified into different groups, characterized by different sets of parameters. The strict stationarity of the GNAR model is established. Two estimation procedures are further developed as well as the asymptotic properties. Numerical studies are conducted to evaluate the finite sample performance of our proposed methodology. At last, two real data examples are presented for illustration purpose. They are the studies of user posting behavior on Sina Weibo platform and air pollution pattern (especially PM 2.5) in mainland China.

EO212 Room P4701 RECENT ADVANCES IN FDR CONTROL METHODOLOGIES Chair: Asaf Weinstein

E0227: **AdaPT: An interactive procedure for multiple testing with side information**
Presenter: **Lihua Lei**, University of California, Berkeley, United States
Co-authors: William Fithian

The problem of multiple hypothesis testing with generic side information is considered: for each hypothesis H_i we observe both a p-value p_i and some predictor x_i encoding contextual information about the hypothesis. For large-scale problems, adaptively focusing power on the more promising hypotheses (those more likely to yield discoveries) can lead to much more powerful multiple testing procedures. We propose a general iterative framework for this problem, called the Adaptive p-value Thresholding (AdaPT) procedure, which adaptively estimates a Bayes-optimal p-value rejection threshold and controls the false discovery rate (FDR) in finite samples. At each iteration of the procedure, the analyst proposes a rejection threshold and observes partially censored p-values, estimates the false discovery proportion (FDP) below the threshold, and proposes another threshold, until the estimated FDP is below. Our procedure is adaptive in an unusually strong sense, permitting the analyst to use any statistical or machine learning method she chooses to estimate the optimal threshold, and to switch between different models at each iteration as information accrues. We demonstrate the favorable performance of AdaPT by comparing it to state-of-the-art methods in various examples.

E0282: **Adaptive filtering procedures for replicability analysis of high-throughput experiments**
Presenter: **Jingshu Wang**, University of Pennsylvania, United States
Co-authors: Weijie Su, Chiara Sabatti, Art Owen

Replicability is a fundamental quality of scientific discoveries. While meta-analysis provides a framework to evaluate the strength of signals across multiple studies accounting for experimental variability, it does not investigate replicability. A single, possibly non-reproducible study, can be enough to bring significance. In contrast, the partial conjunction (PC) alternative hypothesis stipulates that for a chosen number r ($r > 1$), at least r out of n related individual hypotheses are non-null, making it a useful measure of replicability. Motivated by genetics problems, we consider settings where a large number M of partial conjunction null hypotheses are tested, using an $n \times M$ matrix of p-values where n is the number of studies. Applying multiple testing adjustments directly to PC p-values can be very conservative. We here introduce AdaFilter, a new procedure that, mindful of the fact that the PC null is a composite hypothesis, increases power by filtering out unlikely candidate PC hypotheses using the whole p-value matrix. We prove that appropriate versions of AdaFilter control the familywise error rate and the per family error rate under independence. We show that these error rates and the false discovery rate can be controlled under independence and a within-study local dependence structure while achieving much higher power than existing methods. We illustrate the effectiveness of the AdaFilter procedures with three different case studies.

E0341: **Weeding out early false discoveries along the Lasso Path with knockoffs**
Presenter: **Malgorzata Bogdan**, University of Wroclaw, Poland
Co-authors: Weijie Su, Emmanuel Candes, Asaf Weinstein
It is now widely recognized that the popular Lasso method of identfying predictors in large data bases often suffers from including large number of false discoveries. In a recent work this phenomenon has been quantitatively described using the framework of the Approximate Message Passing (AMP) theory. Specifically, it was shown that the Lasso is limited by the FDR-power tradeoff, which in case of moderately dense signals does not allow to simultaneously obtain high power and small false discovery rate. We will use AMP theory to show that this limitation can be overcome by combining Lasso with the recent method of knock-offs, for controlling FDR in the context of multiple regression.

E0551: **A power analysis for knockoffs**
Presenter: **Asaf Weinstein**, Stanford University, United States
Co-authors: Rina Foygel Barber, Emmanuel Candes
Knockoffs is a new framework for controlling the false discovery rate (FDR) in multiple hypothesis testing problems involving complex statistical models. While rigorous results have been obtained regarding type-I error control in a wide range of models, type-II error rates have been far less studied. In general, power calculations are admittedly difficult, in part owing to the very particular structure of the knockoff matrix. Nevertheless, there is a specific setting, involving an i.i.d. Gaussian design, where such calculations are possible. Working in that setting, we leverage recent results to show that a knockoff procedure associated with the Lasso path, achieves close to optimal power with respect to an appropriately defined oracle. This result demonstrates that, in our setting, augmenting the design with fake (knockoff) variables does not have a high cost in terms of power.

EO129 Room P4703 NEW DEVELOPMENTS ON SUFFICIENT DIMENSION REDUCTION Chair: Yichao Wu

E0468: **On some characterizations of, and multidimensional criteria for testing, homogeneity, symmetry and independence**
Presenter: **Feifei Chen**, Renmin University of China, China
Co-authors: Simos Meintanis, Lixing Zhu
Three new characterizations and corresponding distance-based weighted test criteria for the two-sample problem, and for testing symmetry and independence with multivariate data are proposed. All quantities have the common feature of involving characteristic functions, and it is seen that these quantities are intimately related to some earlier methods, thereby generalizing these methods. The connection rests on a special choice of the weight function involved. The new quantities however unlike their predecessors require no moment condition. Equivalent expressions of the distances in terms of densities are given as well as a Bayesian interpretation of the weight function involved. The asymptotic behavior of the tests is investigated both under the null hypothesis and under alternatives. Numerical studies are conducted to examine the performances of the criteria.

E0513: **Transformed variable selection in sufficient dimension reduction**
Presenter: **Yuexiao Dong**, Temple University, United States
Variable transformation with sufficient dimension reduction are combined to achieve model-free variable selection. Existing model-free variable selection methods via sufficient dimension reduction requires a critical assumption that the predictor distribution is elliptically contoured. We suggest a nonparametric variable transformation method after which the predictors become normal. Variable selection is then performed based on the marginally transformed predictors. Asymptotic theory is established to support the proposed method. The desirable variable selection performance of the proposed method is demonstrated through simulation studies and a real data analysis.

E0439: **Efficient estimation in expectile regression using envelope models**
Presenter: **Zhihua Su**, University of Florida, United States
Co-authors: Shanshan Ding, Yi Yang, Tuo Chen
Expectile is an important risk measure with unique advantages and wide applications in the fields of econometrics and finance. The expectile regression (ER) with respect to different expectile levels can provide a comprehensive picture of the conditional distribution of the response variable given the predictors. We adopt an efficient estimation method called the envelope model in ER, and construct a novel envelope expectile regression (EER). Estimation of the EER parameters can be performed using the generalized method of moments (GMM). We establish the consistency and derive the asymptotic distribution of the EER estimators. In addition, we show that the EER estimators are asymptotically more efficient than the ER estimators. Numerical experiments and real data examples are provided to demonstrate the efficiency gains attained by EER compared to ER, and the efficiency gains can further lead to advantages in prediction.

E0657: **On weighted inverse regression ensemble for sufficient dimension reduction and sufficient variable screening**
Presenter: **Zhou Yu**, East China Normal University, China
Based on the conditional characteristic function of the response given the predictors, we introduce weighted inverse regression ensemble (WIRE) as a unified framework for dimension reduction and sufficient variable screening. Unlike classical sufficient dimension reduction estimators and existing sufficient variable selection procedures, WIRE is slicing-free and is readily applicable in the case of multivariate response. Under the setting with fixed predictor dimensionality, the $\sqrt{n}-$consistency of the sample level WIRE estimator is established for dimension reduction. We further propose a forward regression algorithm based on WIRE for ultra-high dimensional feature screening, which enjoys the model-free feature screening consistency when p diverges at an exponential rate of n. The superior finite-sample performances of our proposals over existing methods are demonstrated through extensive simulation studies and the analysis of the Cancel Cell Line Encyclopedia data set.

EO178 Room P4704 COMPUTING IN DESIGN OF EXPERIMENTS Chair: John Stufken

E0274: **Bayesian design for intractable models**
Presenter: **Antony Overstall**, University of Southampton, United Kingdom
Bayesian designs are found by maximising the expectation of a utility function where the utility function is chosen to represent the aim of the experiment. There are several hurdles to overcome when considering Bayesian design for intractable models. Firstly, common to nearly all Bayesian design problems, the expected utility function is not analytically tractable and requires approximation. Secondly, this approximate expected utility needs to be maximised over a potentially high-dimensional design space. To compound these problems, thirdly, the model is intractable, i.e. has no closed form. New approaches to maximise an approximation to the expected utility for intractable models are developed and applied to illustrative exemplar design problems with experimental aims of parameter estimation and model selection.

E0289: **Sequential experiment design for inverse problem from complex dynamic computer codes**
Presenter: **Devon Lin**, Queen's University, Canada
The inverse problem in computer experiments refers to finding an optimal design input that a field observation or a prespecified target as closely as possible. We consider the inverse problem of dynamic computer experiments in which both the field observation and the computer simulator outputs are time series valued. We take a sequential design approach by first fitting a statistical emulator on an initial design set, selecting the

follow-up design points using a novel expected improvement criterion, and updating the emulator. The optimal design input is extracted by the proposed least expected square discrepancy method. The proposed approach is shown to be effective in an illustrative example. Moreover, it achieves better accuracy in extracting the optimal design input compared with existing alternatives in the simulation study and a real application.

E0601: **A construction of cost-efficient designs with guaranteed repeated measurements on interaction effects**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan

A useful class of cost-efficient designs is introduced for two-level multi-factor experiments. Guaranteed repeated measurements on all 2-tuples from any two factors are provided and the number of repetitions is adjusted by the experimenters. Given the number of factors of interest, less resources than an orthogonal array are utilized while its repeated measurement provided a resistance towards outliers that a covering array failed to achieve. To bridge the wide spectrum between two extreme settings (orthogonal arrays and covering arrays) in terms of the number of repeated measures of tuples, we developed a systematic method to construct families of these designs, namely (supersaturated) repeated coverage design, with small run sizes under different number of factors and number of repetitions.

E0258: **Greedy active learning algorithm for logistic regression models**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan

A logistic model-based active learning procedure for binary classification problems is studied, in which we adopt a batch subject selection strategy with a modified sequential experimental design method. Moreover, accompanying the proposed subject selection scheme, we simultaneously conduct a greedy variable selection procedure such that we can update the classification model with all labeled training subjects. The proposed algorithm repeatedly performs both subject and variable selection steps until a prefixed stopping criterion is reached. Our numerical results show that the proposed procedure has competitive performance, with smaller training size and a more compact model, comparing with that of the classifier trained with all variables and a full data set. We also apply the proposed procedure to a well-known wave data set to confirm the performance of our method.

Thursday 21.06.2018 08:30 - 10:10 Parallel Session J – EcoSta2018

EO306 Room B4302 FUNCTIONAL DATA AND COMPLEX STRUCTURES Chair: Hua Liang

E0616: **Quantile estimation for a hybrid model of functional and varying coefficient regressions**
Presenter: **Riquan Zhang**, East China Normal University, China
Co-authors: Jian Zhang, Yanghui Liu, Hui Ding
A hybrid of functional and varying-coefficient regression models for the analysis of mixed functional data is considered. We propose a quantile estimation of this hybrid model as an alternative to the least square approach. Under regularity conditions, we establish the asymptotic normality of the proposed estimator. We show that the estimated slope function can attain the minimax convergence rate as in functional linear regression. A Monte Carlo simulation study and a real data application suggest that the proposed estimation is promising.

E0671: **Scalable and efficient statistical inference for big longitudinal data**
Presenter: **Ling Zhou**, University of Michigan, United States
Co-authors: Peter Song
The theory of statistical inference along with the strategy of divide-and-conquer for large-scale data analysis has recently attracted considerable interest due to great popularity of the MapReduce programming paradigm in the Apache Hadoop software framework. The central analytic task in the development of statistical inference in the MapReduce paradigm pertains to the method of combining results yielded from separately mapped data batches. One seminal solution based on the confidence distribution has recently been established in the setting of maximum likelihood estimation in the literature. The focus is on a more general inferential methodology based on estimating functions, termed as the Rao-type confidence distribution, of which the maximum likelihood is a special case. This generalization provides a unified framework of statistical inference that allows regression analyses of massive data sets of important types in a parallel and scalable fashion via a distributed file system, including longitudinal data analysis, which cannot be handled using the maximum likelihood method. Four important properties of the proposed method are investigated: computational scalability, statistical optimality, methodological generality, and operational robustness. All these properties of the proposed method are illustrated via numerical examples in both simulation studies and real-world data analyses.

E0403: **Likelihood inference for a continuous time GARCH model**
Presenter: **Feng Chen**, UNSW Syd, Australia
Co-authors: Damien Wee, William Dunsmuir
The continuous time GARCH (COGARCH) model is a natural extension of the discrete time GARCH(1,1) model which preserves important features of the GARCH model in the discrete-time setting. However, calibrating the COGARCH model to data is a challenge, especially when observations of the COGARCH process are obtained at irregularly spaced time points. The method of moments has had some success in the case with regularly spaced data, yet it is not clear how to make it work in the more interesting case with irregularly spaced data. As a well-known method of estimation, the maximum likelihood method has not been developed for the COGARCH model, even in the quite simple case with the driving Levy process being compound Poisson, though a quasi-maximum likelihood (QML) method has been proposed. The challenge with the maximum likelihood method in this context is mainly due to the lack of a tractable form for the likelihood. We propose a Monte Carlo method to approximate the likelihood of the compound Poisson driven COGARCH model. We evaluate the performance of the resulting maximum likelihood (ML) estimator using simulated data, and illustrate its application with high frequency exchange rate data.

E0466: **A double application of the Benjamini-Hochberg procedure and its refinement**
Presenter: **Qingyun Cai**, Xiamen University, China
The Benjamini-Hochberg (BH) procedure controls the false discovery rate (FDR), and optimizes signal discovery especially in large dataset. However it does not consider any index information of the null hypotheses. A double application of the BH procedure on two-level hierarchical datasets is proposed. The first application is to identify p-value batches; the second application is to identify null hypotheses rejections in each batch. It is shown that the double application not only maintains the power of BH, but also satisfies an average FDR control and reduces FDR when the signals are clustered. Based on this, a refined procedure of the double application is proposed by incorporating proportion of false null hypotheses to further improve performance.

EO196 Room G4302 ECONOMETRICS OF SPATIAL MODELS, PANELS, AND MODEL UNCERTAINTY Chair: Hon Ho Kwok

E0158: **Network identification methods based on change of basis**
Presenter: **Hon Ho Kwok**, The University of Hong Kong, Hong Kong
Methods for identifying the parameters and networks in linear social interactions models are developed. Three situations are considered: the samples span Rn (long panels); the samples span a proper subspace of Rn (short panels); the models have multiple equilibria. For the short panel situation, the sample vectors are proposed to be represented with respect to a basis of a lower-dimensional space, so that we have fewer regression coefficients, and hence some reduced form submatrices. These submatrices provide equations for identifying the parameters and networks. For the multiple equilibria situation, a method based on matrix triangularization is developed.

E0326: **Autoregressive spectral averaging estimator**
Presenter: **Chu-An Liu**, Academia Sinica, Taiwan
Co-authors: Biing-Shen Kuo, Wen-Jen Tsay
Model averaging in spectral density estimation is considered. We construct the spectral density function by averaging the autoregressive coefficients from all potential autoregressive models and investigate the autoregressive spectral averaging estimator using weights that minimize the Mallows and jackknife criteria. We extend the consistency of the autoregressive spectral estimator to the autoregressive spectral averaging estimator under a condition that imposes a restriction on the relationship between the model weights and autoregressive coefficients. Simulation studies show that the autoregressive spectral averaging estimator compares favorably with the AIC and BIC model selection estimators, and the bias of the averaging estimator approaches zero as the sample size increases.

E0613: **Consistent specification testing under network dependence**
Presenter: **Abhimanyu Gupta**, University of Essex, United Kingdom
Co-authors: Xi Qu
A series-based nonparametric specification test is proposed for a regression function when data are dependent across a network. Our framework permits network dependence to be parametric, parametric with increasing dimension, semiparametric or any combination thereof, thus covering a vast variety of settings. These include spatial error models of varying types and levels of complexity. Despite being applicable so generally, our test statistic is easy to compute and asymptotically standard normal. To prove the latter property, we present a central limit theorem for quadratic forms in linear processes in an increasing dimension setting that may be of independent interest. We also show how our test can be applied to parametric

regression models that become more complex as more data becomes available. Finite sample performance is studied in a simulation study and empirical examples with real data are also included.

E0781: **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 fixed. 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.

EO097 Room G4701 ESTIMATING AND SELECTING MODELS FOR COMPLEX DATA Chair: Garth Tarr

E0161: **Visualising model stability information for better prognosis based network-type feature extraction**
Presenter: **Samuel Mueller**, University of Sydney, Australia
The aim is to present findings to deliver new statistical approaches to identify various types of interpretable feature representations that are prognostically informative in classifying complex diseases. Identifying key features and their regulatory relationships which underlie biological processes is the fundamental objective of much biological research; this includes the study of human disease, with direct and important implications in the development of target therapeutics. We present new and robust ways to visualise valuable information from the thousands of resamples in modern selection methods that use repeated subsampling to identify what features predict best disease progression. We show that using subtractive lack-of-fit measures scales up well to large dimensional situations, making aspects of exhaustive procedures available without its computational cost.

E0266: **More sensitive mixture detection using the empirical moment-generating function**
Presenter: **Michael Stewart**, University of Sydney, Australia
Co-authors: Thomas Porter
The higher criticism method was originally developed as a goodness of fit test to the uniform distribution with focus on multiple testing of a large number of independent p-values. It was shown early on to have very good performance when testing for a certain kind of normal location mixture, indeed it has the same lower-order power properties as a (generalised) likelihood ratio test. It has since been further developed into a tool for feature selection in high-dimensional classification problems and has been shown to have excellent performance in that setting also, both theoretically and according to some well-regarded benchmarking procedures. We provide a higher-order power analysis comparing higher criticism with the generalised likelihood ratio test and another easier-to-implement test based on the empirical moment-generating function, which shows that the latter two tests are optimal in a certain minimax sense whereas higher criticism is not. We also provide some guidance for when to use which method.

E0409: **Efficient semi-parametric generalized linear models based on exponentially tilted splines**
Presenter: **William Aeberhard**, Dalhousie University, Canada
The validity of inference under a generalized linear model directly depends on correctly specifying either a distribution for the response if maximizing a likelihood, or at least a functional form for the variance if resorting to quasi-likelihood and related methods. In order to relax such assumptions, the structure of the exponential family can be exploited to construct semi-parametric models. Indeed, the corresponding probability density function can be seen as an exponential tilt, in the direction of the observation-specific mean, from a common baseline density. The latter can in turn be estimated along with the parameters appearing in the mean equation, which essentially allows to estimate standard errors asymptotically as well as if the true distribution, or variance function, were known. However, so far only computationally prohibitive methods such as empirical likelihood, with minimal bias but maximum variance, have been suggested. We therefore propose an alternative semi-parametric formulation which relies on splines for approximating the logarithm of the baseline density and introduce an efficient algorithm which scales well with the sample size. Simulations and a real data example are presented to illustrate the overall good performances of this alternative method, in particular how it can achieve a usable bias-variance trade off.

E0541: **Local polynomial M-estimation for long memory random design regression models**
Presenter: **Justin Wishart**, Macquarie University, Australia
The local polynomial regression M-estimator for random design regression models that contain long memory structure is explored. The long memory structure is assumed to follow a linear process and the asymptotic behaviour of the local polynomial estimators are established. Time permitting, the numerical performance of the method will be explored on some datasets.

EO085 Room LT-12 HIGH DIMENSIONAL INFERENCE Chair: Huixia Wang

E0160: **Structure adaptive multiple testing**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Jun Chen
Conventional multiple testing procedures often assume that the hypotheses for different units are exchangeable. However, in many scientific applications, external structural information regarding the patterns of signals and nulls are available. We introduce new multiple testing procedures that can incorporate various types of structural information including (partial) ordered structure, smooth structure, group/multi-group structure and covariate-dependent structure. We develop an EM-type algorithm to efficiently implement the proposed procedures and justify the asymptotic validity of our method as the number of hypotheses goes to infinity. We investigate the finite sample performance of the proposed method through extensive simulation studies and real data analysis and find that the new approach is highly competitive to the state-of-the-art approaches in the literature.

E0195: **An adaptive test on high-dimensional parameters in generalized linear models**
Presenter: **Gongjun Xu**, University of Michigan, United States
Significance testing for high-dimensional generalized linear models has been increasingly needed in various applications, however, existing methods are mainly based on a sum of squares of the score vector and only powerful under certain alternative hypotheses. In practice, depending on whether

the true association pattern under an alternative hypothesis is sparse or dense or between, the existing tests may or may not be powerful. We propose an adaptive test that maintains high power across a wide range of scenarios. To evaluate its p-value, its asymptotic null distribution is derived. We conduct simulations to demonstrate the superior performance of the proposed test. In addition, we apply it and other existing tests to an Alzheimer's Disease Neuroimaging Initiative (ADNI) data set, detecting possible associations between Alzheimer's disease and some gene pathways with a large number of single nucleotide polymorphisms (SNPs).

E0515: **Testing high dimensional correlation matrix**
Presenter: **Shurong Zheng**, Northeast Normal University, China
Statistical inferences about sample correlation matrices are of fundamental importance in multivariate analysis, but encounter enormous challenges in the high-dimensional setting. The aim is to test the general structures of high-dimensional correlation matrices. A test statistic is proposed. The limiting null distribution of the test statistic is derived using the random matrix theories. Extensive simulation studies are conducted to demonstrate the finite sample performance of the proposed test. Moreover, two real data examples are provided to show the applicability and the practical utility of the test.

E0519: **MANOVA and change points estimation for high-dimensional longitudinal data**
Presenter: **Ping-Shou Zhong**, Michigan State University, United States
Co-authors: Jun Li, Piotr Kokoszka
The problem of testing temporal homogeneity of p-dimensional population mean vectors from repeated measurements on n subjects over T times is considered. To cope with the challenges brought about by high dimensional longitudinal data, we propose methodology that takes into account not only the "large p, large T and small n" situation, but also the complex temporospatial dependence. We consider both the multivariate analysis of variance (MANOVA) problem and the change point problem. The asymptotic distributions of the proposed test statistics are established under mild conditions. In the change point setting, when the null hypothesis of temporal homogeneity is rejected, we further propose a binary segmentation method and show that it is consistent with a rate that explicitly depends on p, T and n. Simulation studies and an application to fMRI data are provided to demonstrate the performance and applicability of the proposed methods.

EO091 Room LT-13 FRONTIERS IN FINANCIAL STATISTICS Chair: Yichao Wu

E0753: **On the statistical and computational theory for GAN**
Presenter: **Tengyuan Liang**, University of Chicago, United States
The statistical and computational theory for Generative Adversarial Networks (GANs) is studied. On the statistical side, we study the rate of convergence for learning densities under the GANs framework, borrowing insights from nonparametric statistics. We introduce an improved GAN estimator that achieves a faster rate, through leveraging the level of smoothness in the target density and the evaluation metric, which in theory remedies the mode collapse problem reported in the literature. A minimax lower bound is constructed to show that when the dimension is large, the exponent in the rate for the new GAN estimator is near optimal. As a byproduct, we also obtain improved bounds for GAN with deeper ReLU discriminator network. On the computational and algorithmic side, we present a simple yet unified non-asymptotic local convergence theory for smooth two-player games, which subsumes several discrete-time saddle point dynamics. The analysis reveals the surprising nature of the off-diagonal interaction term as both a blessing and a curse. On the one hand, this interaction term explains the origin of the slow-down effect in the convergence of Simultaneous Gradient Ascent (SGA) to stable Nash equilibria. On the other hand, for the unstable equilibria, exponential convergence can be proved thanks to the interaction term, for three modified dynamics which have been proposed to stabilize GAN training. The analysis provides detailed characterization on the choice of learning rate.

E0480: **Volatility of volatility: Estimation and tests based on noisy high frequency data**
Presenter: **Zhiyuan Zhang**, Shanghai University of Finance and Economics, China
Co-authors: Yingying Li, Guangying Liu
A volatility of volatility estimator in a high frequency setting with noise and price jumps is proposed. We establish a feasible central limit theorem for the estimator that has a rate of convergence $n^{1/8}$. To our knowledge, this is the first instance where inference theories for volatility of volatility are obtained under this challenging setup. We further find that the rate of convergence can be improved to $n^{1/5}$ under the null that volatility processes are of bounded variation. This yields a test, which is more powerful than the one based on the general feasible central limit theorem, for the presence of diffusion components in volatility processes. Finite sample performance of the estimator and test statistic are examined by simulation studies. The empirical analysis shows that, for the stocks studied, volatility processes appear to have diffusion components.

E0660: **Statistical learning for optimal personalized wealth management**
Presenter: **Yi Ding**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Yingying Li, Rui Song
A statistical learning method of continuous decision making for investment is proposed. We develop a Q-learning framework that allows one to make personalized wealth management decisions. Statistical properties are established for Q-learning in optimal continuous decision making. As an important application in investment, algorithms for optimal personalized investment decision making are developed. Empirically, we show that the proposed personalized investment decision making rule can substantially improve individuals financial well-being under a framework of consumption based utility analysis.

E0666: **Estimation for high-frequency data under parametric market microstructure noise**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Simon Clinet
A general class of noise-robust estimators is proposed based on the existing estimators in the non-noisy high-frequency data literature. The market microstructure noise is a known parametric function of the limit order book. The noise-robust estimators are constructed as a plug-in version of their counterparts, where we replace the efficient price, which is non-observable in our framework, by an estimator based on the raw price and the limit order book data. We show that the technology can be directly applied to estimate volatility, high-frequency covariance, functionals of volatility and volatility of volatility in a general nonparametric framework where, depending on the problem at hand, price possibly includes infinite jump activity and sampling times encompass asynchronicity and endogeneity.

EO125 Room LT-14 RECENT DEVELOPMENT ON HIGH DIMENSIONAL DATA ANALYSIS Chair: Xiaohui Chang

E0213: **Flexible and efficient estimating equations for variogram estimation**
Presenter: **Xiaohui Chang**, Oregon State University, United States
Variogram estimation plays a vastly important role in spatial modeling. Different methods for variogram estimation can be largely classified into least squares methods and likelihood based methods. A general framework to estimate the variogram through a set of estimating equations is proposed. This approach serves as an alternative approach to likelihood based methods and includes commonly used least squares approaches as its special cases. The proposed method is highly efficient as a low dimensional representation of the weight matrix is employed. The statistical efficiency of various estimators is explored and the lag effect is examined. An application to a hydrology dataset is also presented.

E0570: **Sufficient dimension folding for regressions with matrix- or array-valued predictors**
Presenter: **Wenhui Sheng**, Marquette University, United States
A new sufficient dimension folding method is proposed for regressions in which the predictors are matrix- or array- valued. The method is model-free and avoids strong assumptions on the distribution of X. It does not require kernel or smoothing techniques, neither does it require choosing tuning parameters, such as the number of slices. Moreover, it can deal with both scalar response and multivariate response scenarios. A bootstrap method is introduced to estimate the structural dimension. Asymptotic properties of the estimator is studied. Simulations and real data analysis support the efficiency and effectiveness of the method.

E0789: **Dimension reduction for a stationary multivariate time series**
Presenter: **Chung Eun Lee**, University of Tennessee, Knoxville, United States
Co-authors: Xiaofeng Shao
A new methodology is introduced in order to perform dimension reduction for a stationary multivariate time series. Our method is motivated by the consideration of optimal prediction and focuses on the reduction of the effective dimension in conditional mean of time series given the past information. In particular, we seek a contemporaneous linear transformation such that the transformed time series has two parts with one part being conditionally mean independent of the past. To achieve this goal, we first propose the so-called martingale difference divergence matrix (MDDM), which can quantify the conditional mean independence of $V \in R^p$ given $U \in R^q$ and also encodes the number and form of linear combinations of V that are conditional mean independent of U. Our dimension reduction procedure is based on eigen-decomposition of the cumulative martingale difference divergence matrix, which is an extension of MDDM to the time series context. Some theory is also provided about the rate of convergence of eigenvalue and eigenvector of the sample cumulative MDDM in the fixed-dimensional setting. Favorable finite sample performance is demonstrated via simulations and real data illustrations in comparison with some existing methods.

E0517: **Equality tests of high-dimensional covariance matrices with strongly spiked eigenstructures**
Presenter: **Aki Ishii**, Tokyo University of Science, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima
One of the features of modern data is that the data dimension is extremely high, however, the sample size is relatively low. We call such data HDLSS data. In HDLSS situations, new theories and methodologies are required to develop for statistical inferences. We note that eigenvalues of high-dimensional data grow very rapidly depending on the dimension. There are two types of high-dimensional eigenvalue models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. A lot of works have been done under the NSSE model. We consider equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue (SSE) model. We create a new test procedure on the basis of the high-dimensional eigenstructure. We find the difference of covariance matrices by dividing the high-dimensional eigenstructure into the first eigenspace and the others. We prove that our proposed test procedure has consistency properties both for the size and power. Finally, we check performances of our test procedure in simulations.

EO036 Room LT-15 ADVANCES IN BAYESIAN COMPUTATION Chair: Minh-Ngoc Tran

E0350: **Gaussian variational approximation with a factor covariance structure**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Victor Ong, Michael Smith
Variational approximations have the potential to scale Bayesian computations to large datasets and highly parameterized models. Gaussian approximations are popular, but can be computationally burdensome when an unrestricted covariance matrix is employed and the dimension of the model parameter is high. To circumvent this problem, we consider a factor covariance structure as a parsimonious representation. General stochastic gradient ascent methods are described for efficient implementation, with gradient estimates obtained using the so-called *reparametrization trick*. The end result is a flexible and efficient approach to high-dimensional Gaussian variational approximation. We will illustrate the methodology for robust P-spline regression and some high-dimensional logistic regression models.

E0379: **On Bayesian estimation for the linear ballistic accumulator model**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Gunawan, Thanh Mai Pham Ngoc, Scott Brown
The aim is to estimate a hierarchical version of the Linear Ballistic Accumulator (LBA) model. In general, estimating such models is challenging because the likelihood is an integral over the latent individual random effects and the observations are not Gaussian. Two Bayesian approaches are proposed in order to estimate the hierarchical version of the LBA model. Both methods are based on recent advances in particle Markov chain Monte Carlo (PMCMC)methods. The first approach is an extended version of PMCMC sampler of and the second is based on annealing importance sampling for intractable likelihood (AISIL) method. An estimate of the marginal likelihood is obtained as a by product of the AISIL method. We apply the proposed methods to simulated and real datasets.

E0538: **Gaussian variational approximation for high-dimensional state space models**
Presenter: **Matias Quiroz**, University of New South Wales, Australia
Co-authors: David Nott, Robert Kohn
Variational approximations of the posterior distribution in a high-dimensional state space model are considered. The variational approximation is a multivariate Gaussian density, in which the variational parameters to be optimized are a mean vector and a covariance matrix. The number of parameters in the covariance matrix grows as the square of the number of model parameters, so it is necessary to find simple yet effective parametrizations of the covariance structure when the number of model parameters is large. The joint posterior distribution over the high-dimensional state vectors is approximated using a dynamic factor model, with Markovian dependence in time and a factor covariance structure for the states. This gives a reduced dimension description of the dependence structure for the states, as well as a temporal conditional independence structure similar to that in the true posterior. We illustrate our approach in two high-dimensional applications which are challenging for Markov chain Monte Carlo sampling. The first is a spatio-temporal model for the spread of the Eurasian Collared-Dove across North America. The second is a multivariate stochastic volatility model for financial returns via a Wishart process.

E0432: **Automated sensitivity analysis for Bayesian inference via Markov Chain Monte Carlo**
Presenter: **Dan Zhu**, Monash University, Australia
Co-authors: Liana Jacobi
Bayesian inference relies heavily on numerical Markov chain Monte carlo (MCMC) methods for the estimation of intractable high-dimensional posterior distributions and requires specific inputs. We develop a new general and efficient numerical approach to address important robustness concerns of MCMC analysis with respect to prior input assumptions, a major obstacle to wider acceptance of Bayesian inference, and MCMC algorithm performance (convergence) reflected in dependence to chain starting values. Current input robustness analysis relies heavily on a restrictive and computationally very costly bumping-type approaches based on rerunning the algorithm with a small set of different inputs as well as convergence and efficiency diagnostics based on the autcorrelation of the draws. We introduce a comprehensive input sensitivity analysis based on first order derivatives of MCMC output with respect to the hyper-parameters and starting values to analyse prior robustness and algorithm convergence and efficiency. The approach builds on recent developments in sensitivity analysis of high-dimensional numerical integrals for classical simulation methods using automatic numerical differentiation methods. We introduce a range of new robustness measures to enable researchers to routinely undertake a comprehensive sensitivity analysis of their MCMC results. The methods are implemented for a range of Gibbs samplers and illustrated using both simulated and real data examples.

EO184 Room LT-16 BAYESIAN AND SHRINKAGE ESTIMATION Chair: Antonio Lijoi

E0615: **Shrinkage estimation in the presence of missing data**
Presenter: **Christian Heumann**, Ludwig-Maximilians-University Munich, Germany
Shrinkage estimators such as the LASSO or RIDGE are now popular tools for automatic variable selection in regression analysis. The inference problem has also been studied and various proposals have been made, e.g. Multi-Split-Sampling. An interesting problem is how to get confidence intervals in the presence of missing covariates, when missing is MAR (missing at random) and the data have been multiply imputed. We discuss several approaches to that problem and evaluate the proposed methods in an extensive simulation study.

E0703: **Combined estimation of semiparametric panel data models**
Presenter: **Bai Huang**, Central University of Finance and Economics, China
Co-authors: Tae-Hwy Lee, Aman Ullah
The properties of the combined (model averaging) estimation of semiparametric panel data models with endogeneity are examined. We examine the semiparametric (SP) panel data model with random effect (RE) and fixed effect (FE) and consider a combined estimator of SP RE and SP FE estimators. When the SP RE estimator suffers from inconsistency due to the random individual effect being correlated with the regressors. We show that under certain conditions, the SP combined estimator has strictly smaller risk than SP FE estimator. The asymptotic distribution and risk of the combined estimator are derived using a local asymptotic framework. The Monte Carlo study shows that the SP combined estimator outperforms better than SP FE and SP RE estimators except when the degree of endogeneity or heterogeneity is very small. An empirical application is also presented. According to our calculation of the asymptotic risks of the alternative estimators under comparison, the combined estimation allows researchers to implement efficient estimation under the presence of endogeneity without having to select one of efficient or consistent estimators. Even when there is no endogeneity or when endogeneity is strong, the selection of an efficient estimator or a consistent estimator can be conducted by the combined estimator, as the weights will then be 1 or 0. Hence, the combined estimator is an omnibus estimator across all degrees of endogeneity, especially useful when the endogeneity is weak.

E0758: **Partially linear transformation model for HIV data**
Presenter: **Wei Zhao**, City University of Hong Kong, China
Co-authors: Alan Wan, Peter Gilbert, Yong Zhou
Length-biased and right-censored data arises frequently in practice. A partially linear transformation model is considered for length-biased and right-censored data to account for both the linear and nonlinear covariate effects on survival time. We adopt the local nonlinear technique and develop a global and a local unbiased estimating equations for the simultaneous estimation of unknown covariate effects, which are implemented by an iterative computational algorithm. We establish the asymptotic properties of the proposed estimator under several mild conditions and estimate the standard deviation of the proposed estimator via a bootstrap resampling method. The simulation studies have fully demonstrated the good performance of the proposed estimator under finite sample situation. In addition, the proposed method is further illustrated by two HIV data sets to study the relationship between HIV infection and gender.

E0756: **Bayesian estimation of mean and variance models with penalized splines**
Presenter: **Hector Zarate**, Banco de la Republica de Colombia, Colombia
The fusion among various statistical methods is extended to estimate the mean and variance functions in semiparametric models when the response variable comes from an exponential family distribution. We rely on the natural connection among penalized regression splines that uses basis functions with generalized linear models and Bayesian Markov Chain sampling simulation methodology. The significance and implications of our strategy lie in its potential to contribute to a simple and unified computational methodology that will take into account the factors that affect the variability of the responses which in turn is important for efficient estimation and correct inference for mean parameters without the requirement of fully parametric models. A simulation study investigates the performance of the estimates. Finally, an application to the LIDAR data highlights the merits of our approach.

EO131 Room LT-18 MODERN STATISTICAL METHODS FOR THE COMPLEX DATA Chair: Xingqiu Zhao

E0159: **Variable selection for the random effects two-part model**
Presenter: **Lei Liu**, Washington University in St. Louis, United States
Co-authors: Xiaogang Su, Dongxiao Han, Liuquan Sun
Two-part random effects models have been applied to longitudinal studies for zero-inflated (or semi-continuous) data, characterized by a large portion of zero values and continuous non-zero (positive) values. Examples include monthly medical costs, daily alcohol drinks, etc. As the advance of information technology for data collection and storage, the number of variables available to researchers can be rather large in such studies. To avoid curse of dimensionality and facilitate decision making, it is critically important to select covariates that are truly related to the outcome. We will consider variable selection approaches and apply the "minimum information criterion" method to select variables in the random effects two-part model. The estimation is conducted by adaptive Gaussian quadrature which can be conveniently implemented in SAS Proc NLMIXED. The behavior of our approach is evaluated through simulation, and an application to a longitudinal alcohol dependence study is provided.

E0288: **Weak signals in high-dimension regression: Detection, estimation and prediction**
Presenter: **Hyokyoung Grace Hong**, Michigan State University, United States

Regularization methods, including Lasso, group Lasso and SCAD, typically focus on selecting variables with strong effects while ignoring weak signals. This may result in biased prediction, especially when weak signals outnumber strong signals. The aim is to incorporate weak signals in variable selection, estimation and prediction. We propose a two-stage procedure, consisting of variable selection and post-selection estimation. The variable selection stage involves a covariance-insured screening for detecting weak signals, while the post-selection estimation stage involves a shrinkage estimator for jointly estimating strong and weak signals selected from the first stage. We term the proposed method as the covariance-insured screening based post-selection shrinkage estimator. We establish asymptotic properties for the proposed method and show, via simulations, that incorporating weak signals can improve estimation and prediction performance. We apply the proposed method to predict the annual gross domestic product (GDP) rates based on various socioeconomic indicators for 82 countries.

E0430: **A pooling strategy to effectively use genotype data in quantitative traits genome-wide association studies**
Presenter: **Aiyi Liu**, NIH, United States

The goal of quantitative traits genome-wide association studies (GWAS) is to identify associations between a phenotypic variable, such as a vitamin level, and genetic variants, often single-nucleotide polymorphisms (SNPs). When funding limits the number of assays that can be performed to measure the level of the phenotypic variable, a subgroup of subjects is often randomly selected from the genotype database and the level of the phenotypic variable is then measured for each subject. Because only aproportion of the genotype data can be used, such a simple random sampling method may suffer from substantial loss of efficiency, especially when the number of assays is relative small and the frequency of the less common variant (minor allele frequency) is low. We propose a pooling strategy in which subjects in a randomly selected reference subgroup are aligned with randomly selected subjects from the remaining study subjects to form independent pools; blood samples from subjects in each pool are mixed; and the level of the phenotypic variable is measured for each pool. We demonstrate that the proposed pooling approach produces considerable gains in efficiency over the simple random sampling method for inference concerning the phenotype-genotype association, resulting in higher precision and power. The methods are illustrated using genotypic and phenotypic data from the Trinity Students Study, a quantitative GWAS.

E0659: **Nonparametric comparisons of activity level data from wearable devices**
Presenter: **Hsin-wen Chang**, Academia Sinica, Taiwan
Co-authors: Ian McKeague

The motivation comes from applications to health care monitoring in which there is a need to compare groups of subjects in terms of health outcomes that are functional in nature. We develop nonparametric methods to compare distributions between groups of subjects based on functional data collected from wearable devices. A simulation study shows that the new procedures can deal with unmeasured time-dependent confounders. We illustrate the proposed methods using data from the Darmouth Student Life study.

EO265 Room P4302 CYBERSECURITY RISK MODELING AND PREDICTION Chair: Maochao Xu

E0505: **Functional coefficient additive autoregressive models with measurement error**
Presenter: **Pei Geng**, Illinois State University, United States

Measurement errors are observed in various fields nowadays. For instance, in cyber-security, the records of data breach are reported over time and these observations may be inaccurate due to data collecting techniques. Hence, in the functional coefficient additive autoregressive regression with measurement errors, we propose to apply the local linear estimation method and derive the properties of the proposed estimators in presence of measurement error such as estimation bias and asymptotic distributions. The proposed method is also applied to a data breach example to illustrate the performance.

E0483: **Predicting cyber attacks by deep learning**
Presenter: **Maochao Xu**, Illinios State U, United States

It is important to understand to what extent, and in what perspectives, cyber attacks can be predicted. Being able to predict cyber attacks, even for minutes (if not hours) ahead-of-time, would allow the defender to proactively allocate resources for adequate defense (e.g., dynamically allocating sufficient resources for deep packet inspection or flow-level assembly and analysis). We show that how the deep learning can be used for modeling and predicting the cyber attacks. The real data analysis shows that the proposed deep learning algorithm has a very satisfactory prediction performance.

E0476: **Simultaneous cyber attacks over networks**
Presenter: **Peng Zhao**, Jiangsu Normal University, China

Modeling cyber attacks is a very attractive area of research due to its practical importance. However, most of the related research in the literature does not consider the scenario of simultaneous (or coordinated) attacks which in fact is an important attack instrument in practice. This is mainly because of the complicated evolution of cyber attacks over networks. We propose a novel model which can accommodate different types of simultaneous attacks with possible heterogeneous compromise probabilities. The theoretical results are further validated by the simulation evidence.

E0563: **An efficient algorithm for computing the signatures of networks**
Presenter: **Gaofeng Da**, Nanjing University of Aeronautics and Astronautics, China
Co-authors: Maochao Xu, Ping Shing Ben Chan

Computing the network signature has been an attractive but challenging problem in network reliability. We propose a novel algorithm to compute the signature of a network. This new algorithm relies only on the information of minimal cuts sets or minimal path sets, which is very intuitive and efficient. The new results are used to address the problem of the ageing property of the signature in literature. We further discuss the bounds for the signature when only partial information is available. The application of these new results to cyberattacks is also highlighted.

EO277 Room P4701 ORDER RELATED STATISTICAL INFERENCE Chair: Jong Soo Lee

E0229: **Mean estimate in ranked set sampling using a length-biased concomitant variable**
Presenter: **Tao Li**, Shanghai University of Finance and Economics, China

A ranked set sampling procedure based on the order of a length-biased concomitant variable is proposed. The estimate for population mean based on this sample is given. It is proved that the estimate based on ranked set samples is asymptotically more efficient than the estimate based on simple random samples. The simulation studies are conducted to present the properties of proposed estimate for finite sample size. Moreover, the consequence of ignoring length bias is also address by simulation studies. A real data analysis is discussed at last.

E0239: **Using ranked set sampling with binary outcomes in cluster randomized designs**
Presenter: **Xinlei Wang**, Southern Methodist University, United States
The aim is to study the use of ranked set sampling (RSS) with binary outcomes in cluster randomized designs, where a generalized linear mixed model (GLMM) is used to model the hierarchical data structure involved. Under the GLMM-based framework, we develop different estimators of the treatment effect, including the nonparametric estimator (NP), maximum likelihood estimator (MLE) and pseudo likelihood estimator (PL), and study their properties and performance via numeric evaluation and/or simulation. We also develop procedures to test the existence of the treatment effect based on the three RSS estimators, examine the power and size of the proposed RSS tests, and compare them with existing tests based on simple random sampling (SRS). Further, we illustrate the proposed RSS methods with two data examples, one for rare events and the other for non-rare events. Imperfect ranking is within our consideration. Recommendations are given on whether to use RSS over SRS with binary outcomes in CRDs, and if yes, when to use which RSS estimator among NP, MLE and PL.

E0525: **Unbalancedly sized groups in BRSS-structured cluster randomized designs**
Presenter: **Soohyun Ahn**, Ajou University, Korea, South
In cluster randomized designs (CRDs) structured with balanced ranked set sampling (RSS), we propose a pooled pivotal test to improve the power of detecting any treatment effect by stablizing the estimation of variance components when groups are unevenly sized. We numerically compare the power of the pooled pivotal test with the original pivotal test based on the CRD using balanced RSS and the F-test based on the CRD using simple random sampling (SRS) via simulation. Further, we provide an example using educational data.

E0535: **Bayesian inference for the system lifetimes under Gumbel copulas**
Presenter: **Ping Shing Ben Chan**, The Chinese University of Hong Kong, Hong Kong
The lifetime of a coherent system of n components with identical exponential lifetimes is considered. We derive its density function when the joint distribution of these n components is represented by the Gumbel copulas. Then, the likelihood function of the dependence parameter in the copulas and the rate parameter of the component lifetime based on a random sample of m system lifetimes is constructed. Unfortunately, the likelihood is an unbounded function of the dependence parameter and maximum likelihood estimator does not exist. Therefore we analyze the data via Bayesian inference by assuming the prior distribution of the parameters to be known. The posterior distribution of the unknown parameters is obtained by the Metropolis-Hastings-within-Gibbs algorithm. The proposed method will then be illustrated by a simulated example.

EO273 Room P4703 NON- AND SEMI-PARAMETRIC MIXTURES Chair: Byungtae Seo

E0332: **Doubly smoothed maximum likelihood estimation with application to semiparametric structural measurement error models**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South
The maximum likelihood estimator of structural semiparametric measurement error models is known to be inconsistent when there exist additional error-free covariates. We show that the doubly-smoothed maximum likelihood method can resolve this inconsistency. In addition, we propose a partially doubly-smoothed maximum likelihood method that gives a more efficient estimator than fully smoothed estimators. The universal consistency of the proposed method is discussed along with theoretical and numerical examples.

E0482: **Clustering categorical data using word embedding methods**
Presenter: **Yeojin Chung**, Kookmin University, Korea, South
Clustering continuous data in Euclidean distance has been extensively studied with parametric and nonparametric statistical methods. However, these methods are not directly generalized to categorical data. Particularly clustering for categorical attributes with high cardinality suffers from curse of dimensionality. We propose to convert nominal data into numerical data using word embedding methods such as CBOW or skipgram, which was originally developed for natural language models. With this procedure, each level of the categorical attribute can be represented in a real vector space, where similar (in some sense) categories are located closer. Then well-developed clustering algorithms for continuous data can be used for clustering vectorized categorical data. We compare this approach of clustering categorical data with pre-existing algorithms such as k-medoids or k-modes algorithms.

E0511: **Accelerated failure time modeling via continuous Gaussian scale mixtures**
Presenter: **Sangwook Kang**, Yonsei University, Korea, South
Co-authors: Byungtae Seo
A semiparamtric accelerated failure time (AFT) model resembles the usual linear regression model - the response variable being the logarithm of failure times while the random error term is left unspecified. Thus, it is more flexible than parametric AFT models that assume parametric distributions for the random error term. Estimation for model parameters is typically done through a rank-based procedure, in which the intercept term cannot be estimated. This requires a separate estimation procedure for the intercept, which often leads to unstable estimates. For a better estimation of the intercept essential in estimating mean failure times or survival functions, we propose to employ a mixture model approach. To leave the model as flexible as possible, we consider nonparametric infinite scale mixtures of normal distributions. An expectation-maximization (EM) method is used to estimate model parameters. Finite sample properties of the proposed estimators are investigated via an extensive stimulation study. The proposed estimators are illustrated using a real data analysis.

E0568: **Semiparametric mixture regression with unspecified error distributions**
Presenter: **Weixin Yao**, UC Riverside, United States
In fitting a mixture of linear regression models, normal assumption is traditionally used to model the error and then regression parameters are estimated by the maximum likelihood estimators (MLE). This procedure is not valid if the normal assumption is violated. To relax the assumption on the error distribution hence reduce the modeling bias, we propose semiparametric mixture of linear regression models with unspecified error distributions. We establish a more general identifiability result under weaker conditions than existing results, construct several estimators, and establish their asymptotic properties. These asymptotic results also apply to many existing semiparametric mixture regression estimators whose asymptotic properties have remained unsolved and are considered to be nontrivial to obtain. Using simulation studies, we demonstrate the superiority of the proposed estimators over the MLE when the normal error assumption is violated and the comparability when the error is normal. Analysis of a newly collected Equine Infectious Anemia Virus data in 2017 is employed to illustrate the usefulness of the new results.

EO077 Room P4704 DESIGN OF EXPERIMENTS AND COMPLEX STRUCTURES Chair: Chang-Yun Lin

E0749: **Sequential text-term selection in vector space models**
Presenter: **Feifei Wang**, School of Statistics, Renmin University of China, China
Co-authors: Jingyuan Liu, Hansheng Wang

Text mining has attracted more and more attention with the accumulation of text documents in all fields. We focus on the use of textual information to explain continuous variables in the framework of linear regressions. To handle the unstructured texts, one common practice is to structuralize the text documents via vector space models. However, using words or phrases as the basic analysis terms in vector space models is in high debate. In addition, vector space models often lead to an extremely large term set and suffer from the curse of dimensionality, which makes term selection important and necessary. Toward this end, we propose a novel term screening method for vector space models under a linear regression setup. We first split the whole term space into different subspaces according to the length of terms, and then conduct term screening in a sequential manner. We prove the screening consistency of the method, and assess the empirical performance of the proposed method with simulations based on a dataset of online consumer reviews for cellphones. Then we analyze the associated real data. Results show that the sequential term selection technique can effectively detect the relevant terms by a few steps.

E0792: **Testing covariance matrices in high dimensions**
Presenter: **Danning Li**, Jilin University, China

Testing covariance structure is of significant interest in many areas of high-dimensional inference. Using extreme-value form statistics to test against sparse alternatives and using quadratic form statistics to test against dense alternatives are two important testing procedures. However, quadratic form statistics suffer from low power against sparse alternatives, and extreme-value form statistics suffer from low power against dense alternatives with small disturbances. It would be important and appealing to derive powerful testing procedures against general alternatives (either dense or sparse), which is more realistic in real-world applications. Under the ultra high-dimensional setting, we propose two novel testing procedures with explicit limiting distributions to boost the power against general alternatives.

E0446: **Generalized Bayesian D criterion for single-stratum and multistratum designs**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan

DuMouchel and Jones proposed the Bayesian D criterion by modifying the D-optimality approach to reduce dependence of the selected design on an assumed model. This criterion has been applied to select various single-stratum designs for completely randomized experiments when the number of effects is greater than the sample size. In many industrial experiments, complete randomization is sometimes expensive or infeasible and, hence, designs used for the experiments often have multistratum struc- tures. However, the original Bayesian D criterion was developed under the framework of single-stratum structures and cannot be applied to select multistratum designs. In this paper, we study how to extend the Bayesian approach for more complicated exper- iments and develop the generalized Bayesian D criterion, which generalizes the original Bayesian D criterion and can be applied to select single-stratum and multistratum de- signs for various experiments when the number of effects is greater than the rank of the model matrix.

E0571: **Adaptive design of clinical trials**
Presenter: **Xikui Wang**, University of Manitoba, Canada

The use of adaptive design in Phases I and III clinical trials, particularly following the Bayesian and other approaches, is discussed. The design of Phase I clinical trials is adaptive in nature, and the goal of the design is to search for the maximum tolerated dose while controlling the overall toxicity of all patients in the trial. On the other hand, the goal of response adaptive design of Phase III clinical trials is to maximize the individual ethics while safeguarding the collective ethics. The nature of the adaptive design is the balance between exploration and exploitation according to a certain optimality criterion, which reflects upon the trade-off between the individual ethics of all patients in the trial and the collective ethics of the general public. We will discuss different adaptive designs of clinical trials.

Thursday 21.06.2018 10:40 - 12:20 Parallel Session K – EcoSta2018

EI004 Room LT-18 RECENT DEVELOPMENTS IN HIGH DIMENSIONAL DATA ANALYSIS Chair: Ping-Shou Zhong

E0170: **High-dimensional statistical analysis: Spiked models and data transformation**
Presenter: **Makoto Aoshima**, University of Tsukuba, Japan
Any high-dimensional data is classified into two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. In actual high-dimensional data, a non-sparse and low-rank structure which contains strongly spiked eigenvalues is often found; a structure which fits the SSE model. Under the SSE model, it may be noted that the asymptotic normality of high-dimensional statistics is not valid because it is heavily influenced by strongly spiked eigenvalues. To enable a unified treatment of both the SSE models and non-SSE models, data transformation techniques that transform the SSE models to the non-SSE models were developed previously. Following this novel methodology, strongly spiked eigenvalues are accurately detected by using new PCA-type techniques. With the transformed data, one can create a new statistic which can ensure high accuracy for inferences by using asymptotic normality even under the SSE models. The new techniques to handle high-dimensional data will be demonstrated to solve two-sample problems and classification problems.

E0172: **Community detection of sparse network**
Presenter: **Bing-Yi Jing**, HKUST, Hong Kong
Community detection for networks has been studied intensively in recent years. However, most methods focus on dense networks with little study on sparse networks. We shall investigate ways to detect communities for sparse networks. Simulation results will be given to illustrate the performance of the proposed methods.

E0308: **Order determination for large dimensional matrices**
Presenter: **Lixing Zhu**, Hong Kong Baptist University, Hong Kong
Popularly used eigendecomposition-based criteria such as BIC type, ratio estimation and principal component-based criterion often underestimate model order for regressions or the number of factors for factor models. This longstanding problem is caused by the existence of one or two dominating eigenvalues compared to other nonzero eigenvalues. To alleviate this difficulty, we propose a thresholding double ridge ratio criterion such that the true order can be better identified. Unlike all existing eigendecomposition-based criteria, this criterion can define consistent estimate without requiring the uniqueness of minimum and can then handle possible multiple local minima scenarios. This generic strategy would be readily applied to other dimensionality or order determination problems. We systematically investigate, for general sufficient dimension reduction theory, the dimensionality determination with fixed and divergent dimensions; for local alternative models that converge to its limiting model with fewer projected covariates, discuss when the number of projected covariates can be consistently estimated, when cannot; and for ultra-high dimensional factor models, study the estimation consistency for the number of common factors. Numerical studies are conducted to examine the finite sample performance of the method.

EO093 Room G4701 NEW DEVELOPMENTS IN ANALYZING COMPLEX DATA Chair: Taewook Lee

E0392: **Block wild bootstrap-based CUSUM tests for simultaneous changes of mean and variance robust to high persistence**
Presenter: **Taewook Lee**, Hankuk University of Foreign Studies, Korea, South
Co-authors: Changryong Baek
The conventional CUSUM tests for mean and variance changes tend to be over-sized in the presence of high persistence in mean or/and variance. To overcome such shortcomings, we propose a block wild bootstrap-based CUSUM test (CUSUM-BWB) for simultaneous changes in mean and variance. Our simulation study shows that the CUSUM-BWB tests achieve the correct sizes and comparable powers in finite samples.

E0291: **Robust multilinear rank estimation for tensor regression**
Presenter: **Namgil Lee**, Kangwon National University, Korea, South
Tensor regression refers to a regression analysis whose coefficients and input covariates are in form of multiway arrays, i.e., tensors. The multilinear ranks of a tensor is a generalization of the rank of a matrix in linear algebra into a tensor. We propose a statistical method for robust estimation of multilinear ranks of regression coefficients in tensor regressions assuming tensor-variate generalized linear models (GLMs). A multilinear structure underlying the regression coefficients is shown to cause severe bias in the estimation of the multilinear ranks of higher-order tensors. The proposed method analyzes the multilinear structure in the core tensor obtained from the higher-order singular value decomposition of regression coefficients. Through simulated experiments, it is shown that the proposed method is especially efficient and robust for noisy data and low-rank models, and insensitive to choices of hyperparameters.

E0489: **Sparse smooth backfitting for high-dimensional additive regression**
Presenter: **Eun Ryung Lee**, Sungkyunkwan University, Korea, South
Smooth backtting methods have been proposed and proven as a powerful nonparametric estimation technique for additive regression models in various settings. However, such established studies are restricted to cases with a moderate number of predictors and the existing methods are not directly applicable to high dimensional settings. We develop a new kernel estimator based on smooth backfitting that works in high dimensional additive models. For this, we develop novel penalizations of functional LASSO and its weighted version then they will be applied to smooth backtting methods. We provide oracle results about the resulting penalized smooth backtting methods. In order to implement the new estimators, we derive a numerical algorithm of iteratively applying (componentwise) thresholding operators and present its improved version for a more accurate and efficient computation. Further, we suggest a BIC-type criterion for choosing the penalization parameters.

E0596: **Extended likelihood approach to brain connectivity analysis**
Presenter: **Donghwan Lee**, Ewha Womans University, Korea, South
Co-authors: Youngjo Lee
Conventional multiple testing procedures are commonly used for testing of brain connectivity. However, they are often based on assumptions of independence, so can distort conclusions in brain connectivity analysis. We introduce a hierarchical random effect model for brain connectivity analysis by incorporating a proper correlation structure of test statistics. Based on the extended likelihood approach, we show that the proposed method can provide an accurate estimation of the false discovery rate numerically, and outperforms the other existing methods in terms of validity of error control and power. A real neuroimaging data example for comparing connectivity in two groups are illustrated. We found that an appropriate model is important for the efficiency of connectivity tests.

EO151 Room LT-11 FINANCIAL TIME SERIES ANALYSIS Chair: Yaxing Yang

E0611: **Statistic inference for a single-index ARCH-M model**
Presenter: **Qiang Xiong**, Guangzhou University, China
For a single-index autoregressive conditional heteroscedastic in mean (SI-ARCH-M) model, estimators of the parametric and nonparametric components are proposed by the profile likelihood method. The research results had shown that all the estimators have consistency and asymptotic normality. Based on the asymptotic properties, we propose Wald statistic and generalized likelihood ratio statistic to investigate the testing problems for ARCH effect and goodness of fit, respectively. A simulation study is conducted to evaluate the finite-sample performance of the proposed estimation methodology and testing procedure.

E0614: **Self-weighted LAD-based inference for heavy-tailed continuous threshold autoregressive models**
Presenter: **Yaxing Yang**, Xiamen University, China
The self-weighted least absolute deviation estimation (SLADE) of a heavy-tailed continuous threshold autoregressive (TAR) model is investigated. It is shown that the SLADE is strongly consistent and asymptotically normal. A sign-based portmanteau test is also developed for the diagnostics checking. Simulation studies are carried out to assess the finite-sample performance of our estimator and test. Finally, an empirical example is given to illustrate the usefulness of our method. Combined with a previous work, a complete asymptotic theory on the SLADE of a heavy-tailed TAR model is established. This enriches asymptotic theory on statistical inference for threshold models in the literature.

E0625: **Quasi-likelihood estimation of structure-changed threshold double autoregressive models**
Presenter: **Feifei Guo**, Hong Kong University of Science and Technology, Hong Kong
The quasi-maximum likelihood estimator (QMLE) of the structure-changed and two-regime threshold double autoregressive model is investigated. It is shown that both the estimated threshold and change-point are n-consistent, and they converge weakly to the smallest minimizer of a compound Poisson process and the location of minima of a two-sided random walk, respectively. Other estimated parameters are $n^{(}1/2)$-consistent and asymptotically normal. The performance of the QMLE are assessed via simulation studies and a real example is given to illustrate our procedure.

E0736: **Conditional heteroscedasticity models with time-varying parameters: Estimation and forecasting**
Presenter: **Armin Pourkhanali**, Monash university, Australia
Co-authors: jonathan Keith, Xibin Zhang
The aim is to study the asymptotics and empirical relevance of time-varying generalized autoregressive conditional heteroscedasticity (GARCH) models, where time-varying parameters are approximated by different polynomials of the time variable with unknown orders. In comparison with some existing varying-coefficient GARCH models, our model provides a more flexible mechanism to capture time-varying dynamics of parameters using Chebyshev polynomials. We also investigate the asymptotic properties of these polynomials under mild conditions. Our approach is computationally feasible. The proposed estimation method is justified through Monte Carlo simulation studies and empirical studies. The proposed model is applied to modelling daily returns of the US Treasury bond with a sample period of 30 years, as well as modelling daily returns of the gold futures price. In addition, we compare the out-of-sample forecasting performance of the proposed model with a constant-coefficient GARCH model. The empirical findings support our time-varying GARCH models against its constant-parameter counterpart.

EO216 Room LT-12 SEMIPARAMETRIC METHODS FOR COMPLEX DATA MODELS Chair: Liqun Wang

E0338: **Instrumental variable estimation in ordinal probit models with latent predictors**
Presenter: **Jing Guan**, Tianjin University, China
Co-authors: Liqun Wang
An instrumental variable approach is proposed for the estimation of a probit model with ordinal response and latent predictor variables. We obtain likelihood-based and method of moments estimators which are consistent and asymptotically normally distributed under general conditions. These estimators are easy to compute, perform well and are robust against the normality assumption for the measurement errors in our simulation studies. The proposed method is applied to some real datasets.

E0516: **Observations on bivariate event-times subject to informative censoring**
Presenter: **Dongdong Li**, Simon Fraser University, Canada
Co-authors: Xiaoqiong Joan Hu, John Spinelli
Studies on association between two event-times and how covariates affect the association are often of interest to researchers. Conventional statistical approaches usually assume non-informative censoring which could lead to biased inference when the assumption is violated. We conduct semiparametric regression analysis with right-censored bivariate event-times in presence of informative censoring, using a motivating example which attempts to evaluate how clinical factors (e.g. treatment) affect the risk of getting cardiovascular disease among breast cancer patients who have experienced relapse/second cancer. We propose a pair-wise modeling approach, where the dependence structure between each event-time and the censoring time is modeled through copula functions. We develop a pseudo likelihood-based inference procedure. Simulation studies are conducted to examine the performance of the proposed modelling and inference procedure. Asymptotic properties of the proposed estimator are obtained. The proposed modeling and inference procedure is applied to a motivating research question using breast cancer data for illustration.

E0314: **A goodness-of-fit test for variable-adjusted models**
Presenter: **Chuanlong Xie**, Jinan University, Guangzhou, China
A goodness-of-fit test is provided to checking parametric single-index regression structure when both the response and covariates are measured with distortion errors. Under the null hypothesis, the proposed test statistic asymptotically behaves like a test with univariate covariate and thus, can work better on the significance level maintenance and power performance than existing tests in multivariate covariate cases. With properly selected bandwidths, the proposed test is not seriously affected by distortion measurement errors in the sense that the limiting null distributions in the cases with and without distortion measurement errors can be identical. Numerical studies are conducted to examine the performance of the test in finite sample scenarios.

E0373: **A combined p-value test for the mean difference of high-dimensional data**
Presenter: **Wangli Xu**, Renmin University of China, China
A novel method for testing the equality of high-dimensional means is proposed by employing the idea of multiple hypothesis test. The proposed statistic is the maximum of standardized partial sums of logarithmic p-values. Numerical studies show that the proposed method performs well for both normal and non-normal data. In addition, under both dense and sparse alternative hypotheses, our test can have good power performance. For illustration, a real data analysis is implemented.

EO157 Room LT-14 ADVANCES IN HIGH-DIMENSIONAL AND FUNCTIONAL DATA Chair: Heng Lian

E0268: **Exponential-family random graph models with functional network parameters**
Presenter: **Kevin Lee**, Western Michigan University, United States
Co-authors: Amal Agarwal, Lingzhou Xue
Dynamic networks are a general language for describing time-evolving complex systems, and have long been an interesting research area. It is a fundamental research question to model time varying network parameters. However, due to difficulties in modeling functional network parameters, there is little progress in the current literature to effectively model time varying network parameters. We consider the situation in which network parameters are univariate nonparametric functions instead of constants. Using a kernel regression technique, we introduce a novel unified procedure to effectively estimate those functional network parameters in the exponential-family random graph models. Moreover, by adopting the finite mixture models, we extend our model to mixture of exponential-family random graph models with functional network parameters, which simultaneously allows both modeling and detecting communities for the dynamic networks. The power of our method is demonstrated by simulation studies and real-world applications.

E0592: **Weighted adaptive hard threshold signal approximation**
Presenter: **Xiaoli Gao**, University of North Carolina at Greensboro, United States
The aim is to formulate the copy number into a signal approximation model and to propose a robust change point detection method to simultaneously identify change points and outliers. This proposed method incorporates an individual weight for each observation and adopts the adaptive hard threshold approach to efficiently locate both outliers and copy number variations. The performance of the proposed robust signal approximation method is demonstrated by both simulations and real data analysis. Some theoretical results are also investigated.

E0640: **Sparsity oriented importance learning**
Presenter: **Yi Yang**, McGill University, Canada
Co-authors: Yuhong Yang, Chenglong Ye
With now well-recognized non-negligible model selection uncertainty, data analysts should no longer be satisfied with the output of a single final model from a model selection process, regardless of its sophistication. To improve reliability and reproducibility in model choice, one constructive approach is to make good use of a sound variable importance measure. Although interesting importance measures are available and increasingly used in data analysis, little theoretical justification has been done. We propose a new variable importance measure, sparsity oriented importance learning (SOIL), for high-dimensional regression from a sparse linear modeling perspective by taking into account the variable selection uncertainty via the use of a sensible model weighting. The SOIL method is theoretically shown to have the inclusion/exclusion property: When the model weights are properly around the true model, the SOIL importance can well separate the variables in the true model from the rest. In particular, even if the signal is weak, SOIL rarely gives variables not in the true model significantly higher important values than those in the true model. Extensive simulations in several illustrative settings and real data examples with guided simulations show desirable properties of the SOIL importance in contrast to other importance measures.

E0629: **Uniform knockoff filter for high-dimensional controlled graph recovery**
Presenter: **Jia Zhou**, University of Science and Technology of China, China
Co-authors: Zemin Zheng
Learning the dependence structures in high-dimensional graphical models is of fundamental importance in many contemporary applications. Despite the fast growing literature, procedures with both guaranteed false discovery rate (FDR) control and high power for recovering the graphical structures remain largely unexplored. We develop a new method called uniform knockoff filter that controls the overall FDR in graph recovery based on control variables. Instead of controlling the FDR in a nodewise way, the proposed procedure utilizes a uniform threshold for the statistics based on a large-scale mixture of regression models associated with the graph, which enjoys not only theoretical guarantees of FDR control but also significantly higher power. Furthermore, a scalable implementation approach is developed for the uniform knockoff filter such that all control variables can be generated through a single estimation of the overall graphical structure. Numerical studies verify that our method outperforms existing approaches in power with FDR control.

EO139 Room LT-15 ADVANCES IN HIGH DIMENSIONAL BAYESIAN COMPUTATION Chair: Robert Kohn

E0164: **Hamiltonian Monte Carlo with energy conserving subsampling**
Presenter: **Doan Khue Dung Dang**, University of New South Wales, Australia
Co-authors: Matias Quiroz, Robert Kohn, Minh-Ngoc Tran, Mattias Villani
Hamiltonian Monte Carlo (HMC) has recently received considerable attention in the literature due to its ability to overcome the slow exploration of the parameter space inherent in random walk proposals. In tandem, data subsampling has been extensively used to overcome the computational bottlenecks in posterior sampling algorithms that require evaluating the likelihood over the whole data set, or its gradient. However, while data subsampling has been successful in traditional MCMC algorithms such as Metropolis-Hastings, it has been demonstrated to be unsuccessful in the context of HMC, both in terms of poor sampling efficiency and in producing highly biased inferences. We propose an efficient HMC-within-Gibbs algorithm that utilizes data subsampling to speed up computations and simulates from a slightly perturbed target, which is within $O(m^{(}-2))$ of the true target, where m is the size of the subsample. We also show how to modify the method to obtain exact inference on any function of the parameters. Contrary to previous unsuccessful approaches, we perform subsampling in a way that conserves energy but for a modified Hamiltonian. We can therefore maintain high acceptance rates even for distant proposals. We apply the method for simulating from the posterior distribution of a high-dimensional spline model for bankruptcy data and document speed ups of several orders of magnitude compare to standard HMC and, moreover, demonstrate a negligible bias.

E0234: **Efficient data augmentation techniques for Gaussian state space models**
Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore
A data augmentation scheme is proposed for improving the rate of convergence of the EM algorithm in estimating Gaussian state space models. The scheme is based on a linear transformation of the latent states, and two working parameters are introduced for simultaneous rescaling and re-centering. A variable portion of the mean and scale are thus being moved into the missing data. We derive optimal values of the working parameters (which maximize the speed of the EM algorithm) by minimizing the fraction of missing information. We also study the large sample properties of the working parameters and their dependence on the autocorrelation and signal-to-noise ratio. We show that instant convergence is achievable when the mean is the only unknown parameter and this result is extended to Gibbs samplers and variational Bayes algorithms.

E0380: **Natural gradient factor variational approximations with applications to deep neural network models**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Co-authors: Robert Kohn, David Nott, Nghia Nguyen
Deep neural networks (DNNs) are a powerful tool for functional approximation. We describe flexible versions of generalized linear and generalized

linear mixed models incorporating basis functions formed by a deep neural network. The consideration of neural networks with mixed effects seems little used in the literature, perhaps because of the computational challenges of incorporating subject specific parameters into already complex models. With this in mind, we suggest a computationally efficient variational inference method useful for such models but also applicable more generally. In particular, we develop a natural gradient Gaussian variational approximation method incorporating a factor structure for the covariance matrix which provides a scalable approach to approximate Bayesian inference, even in high dimensions. The use of the natural gradient allows faster and more stable convergence of the variational algorithm, and computation of the natural gradient can be achieved using fast conjugate gradient methods for iterative solution of linear systems, making use of the factor structure of the variational posterior covariance matrix. The proposed methods are illustrated in several examples for DNN random effects models and high-dimensional logistic regression with sparse signal shrinkage priors.

E0706: **The Block-Poisson estimator for efficient Bayesian inference in intractable models**
Presenter: **Mattias Villani**, Linkoping University, Sweden
Co-authors: Matias Quiroz, Robert Kohn, Minh-Ngoc Tran, Doan Khue Dung Dang
Many applications involve models with intractable likelihood functions, for example random effects models, big data problems with costly likelihood evaluations, and the class of doubly intractable problems where the likelihood contains an intractable normalization constant. We propose an efficient dependent pseudo-marginal Markov Chain Monte Carlo algorithm based on the Block-Poisson estimator of the intractable likelihood. The Block-Poisson estimator is shown to be unbiased and provides a convenient way to introduce dependence between successive likelihood estimates, which is well known to be crucial for pseudo-marginal algorithms to work well with noisy estimates. The Block-Poisson estimator can give occasionally negative estimates and we follow previous research and run our MCMC on the absolute value of the posterior density followed by an importance sampling correction to obtain simulation consistent estimates of posterior expectations of functions. We provide guidelines for optimally tuning the algorithm that balances the computational cost of the likelihood estimator against the loss of efficiency from using noisy estimates and the importance sampling inefficiency resulting from occasional negative estimates. We demonstrate the method on subsampling in big data problems and show dramatical speed-ups compared to regular MCMC and another recently proposed exact subsampling algorithm.

EO326 Room LT-16 ECONOMICS/STATISTICS METHODS IN BIOMEDICAL RESEARCH Chair: Shouhao Zhou

E0684: **Understand cancer natural history from cancer screening trials**
Presenter: **Yu Shen**, UT MD Anderson Cancer Center, United States
The primary objective of screening is to detect a cancer in the preclinical state with the expectation that earlier diagnosis together with effective treatment will yield longer survival or cure. Over last decades, many of cancer screening studies have been conducted for breast, lung, colon and prostate. These trials generate data, which may be used to estimate the preclinical sojourn time distribution and screening sensitivity and other quantities of interest from the screening cohort. The information on the natural history of cancer is critical in designing optimal screening programs and assessing screening benefit. We will review some existing parametric and nonparametric approaches in this area.

E0674: **A Bayesian screening approach for hepatocellular carcinoma using two longitudinal biomarkers**
Presenter: **Nabihah Tayob**, The University of Texas MD Anderson Cancer Center, United States
Co-authors: Francesco Stingo, Kim-Anh Do, Ziding Feng, Anna Lok
Advanced hepatocellular carcinoma (HCC) has limited treatment options and poor survival. Early detection of HCC is critical to improve the prognosis of these patients. Current guidelines for high-risk patients include six-month ultrasound screenings but these are not sensitive for early HCC. Alpha-fetoprotein (AFP) is a widely used diagnostic biomarker but has shown limited use in HCC screening with a fixed threshold. Approaches that incorporate longitudinal AFP have shown potentially increased detection of HCC however AFP is not elevated in all HCC cases. We incorporate a second HCC biomarker, des-gamma-carboxy prothrombin (DCP). The Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) Trial is a valuable source of data to study biomarker screening. We assume the trajectories of AFP and DCP follow a joint hierarchical mixture model with random change points. Markov chain Monte Carlo methods are used to estimate the posterior distributions used in risk calculations among future patients. The posterior risk of HCC, given longitudinal values of AFP and DCP, is used to determine whether a patient has a positive screen. The screening algorithm was compared to alternatives in the HALT-C Trial (using cross-validation) and in simulations studies under a variety of possible scenarios.

E0603: **A Bayesian dose-finding design in oncology using pharmacokinetic/pharmacodynamic modeling**
Presenter: **Yisheng Li**, The University of Texas MD Anderson Cancer Center, United States
Co-authors: Xiao Su, Peter Mueller, Kim-Anh Do
Most oncology phase I dose-finding trial designs are either algorithmic or empirical model based, with the goal of identifying the maximum tolerated dose (MTD). Efficiency may be lost in these designs due to failure to exploit the drug concentration-time profile that can be modeled based on available pharmacokinetic (PK) data. To extend the existing designs that make use of low-dimensional summaries of the PK profiles and/or PD endpoints, we propose a Bayesian framework to model the dose-concentration-pharmacologic effect-toxicity (D-C-E-C) relationship, by using PK/PD modeling and modeling of the relationship between a latent PD outcome (dynamic pharmacologic effect) and a binary toxicity endpoint. An important feature of the proposed framework is that it allows explicit modeling of the effects of treatment schedule and method of administration (e.g., drug formulation and route of administration), which are difficult to model under the existing design framework. We compare the performance of the proposed designs with common phase I designs and a nonparametric benchmark design via simulation studies. We illustrate the proposed design by applying it to a phase I trial of a γ-secretase inhibitor in metastatic or locally advanced solid tumors conducted at MD Anderson Cancer Center.

E0649: **A semiparametric approach to model Medicare insurance choice and expenditures**
Presenter: **Chan Shen**, University of Texas MD Anderson Cancer Center, United States
Medicare insurance covers a large percentage of the US population. Within Medicare, there are several choices beneficiaries can make including whether to purchase supplementary insurance and whether to choose Health Management Organization (HMO) plans. There are debates about the impact of these choices on medical expenses. We use a semiparametric approach to model Medicare insurance choice and expenditures. For insurance choice, this semiparametric approach allows a very flexible form for choice probabilities, thereby avoiding the independence of irrelevance alternatives assumption in multinomial logit models. In addition to a number of important demographic variables, we find that that market penetration of HMO, healthcare knowledge level and number of chronic conditions have significant impacts on insurance choice. For health expenditures, we treat the insurance variables as endogenous choices whose impacts flexibly depend on exogenous covariates. Both equations are estimated using a recursive-differencing estimator for a nonparametric expectation, which has desirable asymptotic and finite sample properties. We used the Medicare Beneficiary Survey data to create the study cohort.

EO190 Room P4302 ADVANCES IN REGRESSION AND NETWORK DATA ANALYSIS Chair: Binyan Jiang

E0396: **On cumulative slicing estimation for high dimensional data**
Presenter: **Cheng Wang**, Shanghai Jiao Tong University, China
In the context of sufficient dimension reduction (SDR), sliced inverse regression (SIR) is the first and perhaps one of the most popular tools to reduce the covariate dimension for high dimensional nonlinear regressions. Despite the fact that the performance of SIR is very insensitive to the number of slices when the covariate is low or moderate dimensional, our empirical studies indicate that, the performance of SIR relies heavily upon the number of slices when the covariate is high or ultrahigh dimensional. How to select the optimal number of slices for SIR is still a longstanding problem in the SDR literature, which is a crucial issue for SIR to be effective in high and ultrahigh dimensional regressions. We consider an improved version of SIR, the cumulative slicing estimation (CUME) method, which does not require selecting the optimal number of slices. We provide a general framework to analyze the phase transition phenomenon for the CUME method. We show that, without sparsity assumption, CUME is consistent if and only if $p/n \to 0$, where p stands for the covariate dimension and n stands for the sample size. If we make certain sparsity assumptions, then the thresholding estimate for the CUME method is consistent as long as $log(p)/n \to 0$. We demonstrate the superior performance of our proposals through extensive numerical experiments.

E0393: **Undirected network models with degree heterogeneity and homophily**
Presenter: **Ting Yan**, Central China Normal University, China
The degree heterogeneity and homophily are two typical features in network data. We formulate a general model for undirected networks with these two features and present the moment estimation for inferring the degree parameter and homophily parameter. We establish a unified theoretical framework for the moment estimator. In particular, we establish conditions under which the consistency and asymptotic normality of the estimator hold. We apply it to some special cases. Numerical studies demonstrate our theoretical findings.

E0414: **Finite sample goodness-of-fit tests for the stochastic block model**
Presenter: **Vishesh Karwa**, The Ohio State University, United States
Stochastic Block models (SBM) with unknown block structure are widely used to detect community structure in real world networks. SBM comes in many variants, hence it is essential to evaluate the fit of these variants. Testing the goodness-of-fit of such models is a challenging task due to the fact that the parameters of an SBM are usually estimated from a single observed network. Usual asymptotic tests are not valid. We will introduce three different variations of Stochastic Block Models and present a finite sample goodness-of-fit test for these models, when the block structure is unknown. The finite sample test is based on extending the classic fisher's exact test to SBMs with known and unknown blocks. The machinery of Algebraic Statistics is used in constructing this test. In particular, a key building block for the test is a sampler from the so called "fibers" of SBMs with known block assignments - the set of all graphs with a fixed sufficient statistic. Sampling from these fibers is carried out using Markov bases.

E0676: **Multi-connection selection and estimation**
Presenter: **Fengrong Wei**, University of West Georgia, United States
Network analysis becomes popular nowadays and subjects within a network could be complexly connected through multiple relationships. A network of N nodes with p types of connections is investigated and a network regression approach to investigate the heterogeneous impacts of these connections on a continuous response is proposed. The structure of each connection is taken into consideration in the proposed approach to select the important connections and estimate their corresponding effects on the response simultaneously. Simulation studies demonstrate the effectiveness of the proposed approach in both connection selection and heterogeneous effect estimation. The usefulness of the proposed approach is further illustrated in a real example. Extension studies are discussed and the corresponding implementable methodologies are provided.

EO046 Room P4701 NEW COMPUTATIONAL METHODS FOR STATISTICAL INFERENCE Chair: Dungang Liu

E0662: **Semi-supervised inference for explained variance in high-dimensional linear regression and its applications**
Presenter: **Zijian Guo**, Rutgers University, United States
Statistical inference is considered for the explained variance under the high-dimensional linear model in the semi-supervised setting. A calibrated estimator, which efficiently integrates both labelled and unlabelled data, is proposed. It is shown that the estimator achieves the minimax optimal rate of convergence in the general semi-supervised framework. The optimality result characterizes how the unlabelled data affects the minimax optimal rate. Moreover, the limiting distribution for the proposed estimator is established and data-driven confidence intervals for the explained variance are constructed. We further develop a randomized calibration technique for statistical inference in the presence of weak signals and apply the obtained inference results to a range of important statistical problems, including signal detection and global testing, prediction accuracy evaluation, and confidence ball construction. The numerical performance of the proposed methodology is demonstrated in simulation studies and an analysis of estimating heritability for a yeast segregant data set with multiple traits.

E0704: **Modeling hybrid traits for comorbidity**
Presenter: **Heping Zhang**, Yale University, United States
Co-authors: Dungang Liu, Jiwei Zhao, Xuan Bi
A novel multivariate model for analyzing hybrid traits and identifying genetic factors for comorbid conditions. Comorbidity is common phenomenon in mental health that an individual suffers from multiple disorders simultaneously. In the Study of Addiction: Genetics and Environment (SAGE), alcohol and nicotine addiction were recorded through multiple assessments that we refer to as hybrid traits. Statistical inference for studying the genetic basis of hybrid traits has not been well-developed. Rank-based methods do not inform the strength or direction of effects. Parametric frameworks have been proposed in theory, but they are neither well-developed nor extensively used in practice due to their reliance on complicated likelihood functions that have high computational complexity. Many existing parametric frameworks tend to instead use pseudo-likelihoods to reduce computational burdens. We develop a model fitting algorithm for the full likelihood. Our simulation studies demonstrate that inference based on the full likelihood can control the type-I error rate, and gains power and improves the effect size estimation when compared with several existing methods. These advantages remain even if the distribution of the latent variables is misspecified. For the SAGE data, we identify three genetic variants that are significantly associated with the comorbidity of alcohol and nicotine addiction at the chromosome-wide level.

E0564: **Data linking approaches for meta-analysis of individual participant data**
Presenter: **EY Mun**, University of North Texas Health Science Center, United States
Clinical trials are heterogeneous in key design features, including participants, treatments, comparisons, outcome measures, and settings, creating challenges for feasibility and interpretation for complex multivariate research synthesis. Such between-study heterogeneity has posed a significant barrier to fully utilizing individual participant data for meta-analysis applications despite well-known advantages of analyzing individual participant data in meta-analysis. We present several methodological approaches we have adopted to address between-study heterogeneity for Project INTEGRATE, a large-scale research synthesis project utilizing individual participant data, as well as aggregate data, from multiple independent trials that were developed to prevent alcohol misuse for adolescents and college students. Of the approaches taken for Project INTEGRATE to

link data validly across samples and studies, we will focus on a mapping approach and a Bayesian multilevel modeling approach and present data application examples.

E0707: **Optimal and adaptive P-value combination methods**
Presenter: **Zheyang Wu**, Worcester Polytechnic Institute, United States
P-value combination is an important statistical approach for information-aggregated decision making. It is foundational to a lot of applications such as meta-analysis, data integration, signal detection, and others. We propose two generic statistic families for combining p-values: gGOF, a general family of goodness-of-fit type statistics, and tFisher, a family of Fisher type p-value combination with a general weighting-and-truncation scheme. The two families unify many optimal statistics over a wide spectrum of signal patterns, including both sparse and dense signals. We provide efficient solutions for analytical calculations of p-value and statistical power, as well as studies of asymptotic efficiencies, while emphasizing the conditions of realistic data analysis: small or moderate group size, Gaussian and non-Gaussian distributions, correlated data, generalized linear model based alternative hypotheses, etc. Based on these two families of statistics, omnibus tests are also designed for adapting the family-retained advantages to unknown signal patterns. Applications of these methods are illustrated in analyses of large-scale omics data.

EO107 Room P4703 MODELLING AND CLUSTERING METHODS FOR ANALYZING COMPLEX PROCESSES Chair: Xiaoling Dou

E0298: **Analysis of a disaster prevention consciousness survey using a small area model based approach**
Presenter: **Masayo Hirose**, Institute of Statistical Mathematics, Japan
Co-authors: Yoosung Park, Takahiro Tsuchiya
There had been conducted a mail survey in order to determine the disaster prevention consciousness of the residents of a particular city in Japan. We suggest that ascertaining such consciousness of residents of smaller administrative divisions would be even more useful, in order to help plan more suitable services for the residents. However, there are concerns that such subdividing could reduce the sample size within each division, and the conventional approach used in public administration research might yield unreasonable estimates of consciousness level. We thus adopted a model based approach to analyze such survey data, and compared the results with results obtained using the conventional approach. We found that the model-based approach dramatically reduced several large differences observed between estimates, originally provided by the conventional approach, among each neighborhood. Although the model based approach is well known and used frequently in the field of small area estimation, to the best of our knowledge, this is the first study in Japan to adopt this approach for an analysis of a consciousness survey at the municipality level.

E0540: **On trend change mechanisms of financial markets**
Presenter: **Yoko Tanokura**, Meiji University, Japan
Co-authors: Sato Seisho, Genshiro Kitagawa
Political events and economic problems in one country have increasingly influenced the economies and financial markets worldwide. For example, the global economic crisis occurred at the end of 2008 was first triggered by the US subprime mortgage problem in the summer of 2007. In a financial market, the trend of the asset price can be regarded as the gradually changing long-term fluctuations caused by characteristics specific to the asset. On the other hand, the short-term cyclical fluctuations can sensitively be influenced by those of any other asset prices, and may lead to a future change in the trend direction. Aiming to analyze the mechanism of causing changes of the long-term trend, we extract the long-term trend component, the seasonal component if exists, and the short-term cyclical component from a financial market index, based on a seasonal adjustment model. The aim is to investigate the fluctuation structure of the long-term trend component of the Nikkei average, focusing on the relationship with the short-term cyclical components. It is found that the short-term cyclical component of the Nikkei average which was influenced by those of the other stock market indices, prompted changes of the long-term trend component of the Nikkei average.

E0610: **Clustering models for earthquake occurrences and extensions**
Presenter: **Jiancang Zhuang**, Institute of Statistical Mathematics, Japan
The Epidemic-Type Aftershock Sequence (ETAS) model has become a de facto standard model, or null hypotheses, for testing other models and hypotheses related to seismicity. The purpose is to summarize the history of the ETAS model formulation and some other topics, including 1. Stochastic separation of earthquake clusters from the catalog; 2. Modelling the role of earthquake fault geometry in seismicity triggering and inversion of earthquake fault geometry based on seismicity; and 3. Incorporating depth and focal mechanism into the model formulation. The focus is on the methods of finding the features that are in the earthquake data but not described in the model. Such methods can also be applied to other point process models.

E0169: **A nonparametric functional clustering of mouse ultrasonic vocalization data**
Presenter: **Xiaoling Dou**, Waseda University, Japan
A nonparametric method of functional clustering is proposed. The method is based on functional principle component analysis. We calculate the principle component score for each functional data, and estimate the density of the principle component score. Then we find the mode for each principle component, and classify the functional data by calculating the distance of the functional data from the mode. A real dataset of mouse ultrasonic vocalization is illustrated.

EO153 Room P4704 STATISTICAL MODELLING AND INFERENCE IN DIRECTIONAL STATISTICS Chair: Toshihiro Abe

E0413: **The mixture transition distribution modeling for higher order circular Markov processes**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan
Co-authors: Takayuki Shiohama
The stationary higher order Markov process for circular data is considered. We employ the mixture transition distribution model to express the transition density of the process. The underlying circular transition distribution is based on the Wehrly and Johnson's bivariate circular models. The structure of the circular autocorrelation function is found to be similar to the autocorrelation function of the AR process on the line. The validity of the model is assessed by applying it to a series of real directional data.

E0490: **Models for cylindrical data and their applications**
Presenter: **Toshihiro Abe**, Nanzan University, Japan
Recent cylindrical models are reviewed. The WeiSSVM model has numerous good properties, such as simple normalizing constant and hence very tractable density, parameter-parsimony and interpretability, maximum entropy characterization, good circular-linear dependence structure, easy random number generation thanks to known marginal/conditional distributions, and flexibility illustrated via excellent fitting abilities. We also review some other related models. As an illustrative example, some of the models are applied in analyses of cylindrical data set.

E0497: **A cylindrical distribution whose linear part is heavy-tailed**

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

Co-authors: Kunio Shimizu, Toshihiro Abe

There exist many examples of heavy-tailed phenomena such as insurance losses and returns in financial data, heavy-precipitation data, and heavy burst of teletransmission and Internet activity. Potential applications are combinations of linear and circular data through 24-hours clock for example as the circular part. If the cylindrical distributions whose linear parts can model only light-tailedness are applied to such data, the estimation and test may be biased by linear large observations, and it leads to wrong results. We propose a cylindrical distribution heavy-tailed for the linear part through a generalized Gamma mixture of Abe-Ley distribution whose linear part is related to a Weibull distribution. The conditional distribution of the linear variable given circular variable is a generalized Pareto distribution and therefore, it might not have any conditional moments, but the mode and median are expressed by closed forms. As an illustrative example, we fit the proposed distribution with likelihood techniques to earthquake data, which consists of the turning angles for epicenters and magnitude during 72 hours before the 2011 Great East Japan Earthquake, and compare the result with those by other cylindrical distributions each of whose linear part model only light-tailedness.

E0602: **Circular time series analysis based on the projected normal distribution**

Presenter: **Takayuki Shiohama**, Tokyo University of Science, Japan

Co-authors: Takuto Kotsubo, Hiroaki Ogata

A new approach to a circular time series modeling is introduced which is based on the projected normal distributions with circular-circular regression. Some new perspectives on circular partial autocorrelation coefficients are introduced and its statistical inferences are discussed. Maximum likelihood estimation for the unknown model parameters and its asymptotic properties are investigated. Numerical simulations are provided to demonstrate the performance of the proposed approach. The resulting models are used to illustrate the higher order dependency of the wind direction time series.

Thursday 21.06.2018 13:50 - 15:30 Parallel Session L – EcoSta2018

EO164 Room G4701 RECENT ADVANCES AND CHALLENGES IN HIGH DIMENSIONAL DATA Chair: Yuan Ke

E0775: **An alternative ADMM algorithm for computing the graph-fused lasso**
Presenter: **Teng Zhang**, University of Central Florida, United States
Co-authors: Yi Yang
An algorithm for computing the graph fused lasso is introduced which is based on a special method of applying the alternating direction method of multipliers (ADMM). Compared to the standard ADMM algorithm, the proposed algorithm applies a different way of decomposing the objective function into several parts, which improves the computational complexity per iteration from $O(n^2)$ to $O(n)$, where n is the number of samples. The principle of this algorithm can also be applied to other problems such as trend filtering.

E0696: **Multiple change point detection for manifold-valued data with applications to dynamic functional connectivity**
Presenter: **Qiang Sun**, University of Toronto, United States
In neuroscience, functional connectivity describes the connectivity between brain regions that share functional properties. It is often characterized by a time series of covariance matrices between functional measurements of distributed neuron areas. An effective statistical model for functional connectivity and its changes over time is critical for better understanding neurological diseases. To this end, we propose a log-mean model with an additive heterogeneous noise for modeling random symmetric positive definite matrices that lie in a Riemannian manifold. The heterogeneity of error terms is introduced specifically to capture the curved nature of the manifold. A scan statistic is then developed for the purpose of multiple change point detection. Despite that the proposed model is linear and additive, it is able to account for the curved nature of the Riemannian manifold. Theoretically, we establish the sure coverage property. Simulation studies and an application to the Human Connectome Project lend further support to the proposed methodology.

E0721: **Fast convergence of Newton-type methods on high-dimensional problems**
Presenter: **Yuekai Sun**, University of Michigan, United States
The convergence rate of Newton-type methods on high-dimensional problems is studied. The high-dimensional nature of the problem precludes the usual global strong convexity and smoothness that underlie the classical analysis of such methods. We find that restricted version of these conditions which typically arise in the study of the statistical properties of the solutions are also enough to ensure good computational properties of Newton-type methods. We explore the algorithmic consequences in distributed and online settings.

E0294: **On testing for high dimensional white noise**
Presenter: **Zeng Li**, Pennsylvania State University, United States
Testing for white noise is a classical yet important problem in statistics, especially for diagnostic checks in time series modeling and linear regression. For high-dimensional time series in the sense that the dimension p is large in relation to the sample size T, the popular omnibus tests including the multivariate Hosking and Li-McLeod tests are extremely conservative, leading to substantial power loss. To develop more relevant tests for high-dimensional cases, we propose a portmanteau-type test statistic which is the sum of squared singular values of the first q lagged sample autocovariance matrices. Using the tools from random matrix theory and assuming both p and T diverge to infinity, we derive the asymptotic normality of the test statistic under both the null and a specific VMA(1) alternative hypothesis. Non-trivial estimations are proposed for these parameters and their integration leads to a practically usable test. Extensive simulation confirms the excellent finite-sample performance of the new test with accurate size and satisfactory power for a large range of finite (p,T) combinations, therefore ensuring wide applicability in practice. In particular, the new tests are consistently superior to the traditional Hosking and Li-McLeod tests.

EO245 Room LT-11 COPULAS AND DEPENDENCE IN ECONOMETRICS AND STATISTICS Chair: Hao Wang

E0165: **Time-varying extreme value dependence with application to leading European stock markets**
Presenter: **Daniela Castro-Camilo**, King Abdullah University of Science and Technology, Saudi Arabia
Extremal dependence between international stock markets is of particular interest in today's global financial landscape. However, previous studies have shown this dependence is not necessarily stationary over time. We concern ourselves with modeling extreme value dependence when that dependence is changing over time, or other suitable covariate. Working within a framework of asymptotic dependence, we introduce a regression model for the angular density of a bivariate extreme value distribution that allows us to assess how extremal dependence evolves over a covariate. We apply the proposed model to assess the dynamics governing extremal dependence of some leading European stock markets over the last three decades, and find evidence of an increase in extremal dependence over recent years.

E0264: **Bayesian nonparametric inference via skew-t copulas**
Presenter: **Xue Wang**, University of Kent at Canterbury, United Kingdom
A copula can fully characterize the dependence of multiple variables. The aim is to present a general Bayesian nonparametric approach to the modeling and estimation of the copula density function. The idea is to use infinite mixture models, based on a family of skew-t copulas, to construct such a flexible family of copula densities. We first introduce the skew-t copula, which we then extend to an infinite mixture model. An MCMC algorithm is developed to draw samples from the correct posterior distribution and the model is investigated using both simulated and real applications.

E0524: **Portfolio diversification strategy via ARMA-GARCH vine copula approach**
Presenter: **Hao Ji**, Northwest Agricultural and Forestry University, China
Under the framework of the Black-Litterman model, the copula-opinion pooling method (COP method) does not consider heteroscedasticity and autocorrelation of assets, and is not flexible enough to describe the dependence structure compared with the vine copula in high-dimensional cases. A portfolio diversification application is extended to the ARMA-GARCH-Vine-Copula approach in order to overcome the drawbacks of COP method. First, the ARMA-GARCH model is used for each univariate returns series to remove the heteroscedasticity and autocorrelation. Then, conditional Spearman's has been chosen as the measurement to classify clusters of candidate assets. Algorithms for the numerical estimation of conditional Spearman's are also illustrated. After performing the cluster algorithms, a vine copula is used to model the dependence structure among assets returns. We choose one asset from each cluster as a candidate asset to avoid co-moving in their lower regions. Finally, a strategy is presented to construct diversified portfolios at one day forecast horizon by adding the investors information into portfolio selection procedure and minimizing the CVaR. It can be considered as an extension of the Black-Litterman model. The presented methodology is expected to be useful for the selection of a diversified portfolio of asset returns.

E0646: **Full-range tail dependence copulas with insurance applications**
Presenter: **Jianxi Su**, Purdue University, United States
Co-authors: Lei Hua
Copulas are an important tool to formulating models for multivariate data analysis. An ideal copula should conform to a wide range of problems at hand, being either symmetric or asymmetric, and exhibiting flexible extent of tail dependence. The copula to be discussed is exactly one such candidate. Specifically, a class of full-range tail dependence copulas will be introduced which has been proved quite useful for modeling dependent (insurance/financial) data. The key mechanism for constructing such flexible copula models and some future research related to this topic will be discussed.

EO283 Room LT-12 ANALYSIS OF BIG DATA: AN INTEGRATION PERSPECTIVE Chair: Kin Yat Liu

E0241: **Integrative sparse principal component analysis of multiple heterogeneous datasets**
Presenter: **Kuangnan Fang**, Xiamen University, China
With high-dimensional covariates and a small sample size, the analysis of a single dataset often generates unsatisfactory results. In a series of studies, it is shown that the integrative analysis of multiple independent datasets provides an effective way of pooling information and outperforms single-dataset analysis and may alternative multi-datasets analyses, especially including the classic meta-analysis. Compared to regression analysis + variable selection, integrative analysis has not been well conducted based on dimension reduction techniques. We conduct the integrative analysis of multiple heterogeneous datasets based on the sparse principal component analysis (SPCA) technique. A penalization approach is adopted for regularized estimation and selection of important loadings. Significantly advancing from the existing integrative analysis studies, we take advantage of the similarity across datasets and impose contrasted penalties to generate more accurate estimation/selection. Multiple similarity conditions are comprehensively considered. Statistical properties of the proposed iSPCA (integrative SPCA) approach are established, and effective computational algorithms are developed. A wide spectrum of simulations demonstrate competitive performance of iSPCA over the alternatives. Two sets of data analysis further establish its practical applicability.

E0283: **Identifying the subpopulation-specific covariates in FMR model**
Presenter: **Mengque Liu**, The School of Economics, Xiamen University, China
A finite mixture of regression (FMR) models for high dimensional inhomogeneous data is considered, where the number of covariates may be much larger than sample size. However, there lack the mechanism to analyze the sub-population characterisation. We propose an l_0 norm penalty which is the first to identify the subpopulation-specific important covariates in FMR models. Computationally it is realized using an efficient EM algorithm. Theoretically it has the much desired consistency properties, its oracle results are also provided. Simulation study under diverse settings shows the superior performance of the proposed method. In both simulations and real data analysis, we demonstrate a significant gain in identification rate over the FMRLssso method.

E0383: **A joint learning of multiple precision matrices with sign consistency**
Presenter: **Yuan Huang**, University of Iowa, United States
The Gaussian graphical model is a popular tool for inferring the relationships among random variables, where the precision matrix has a natural interpretation of conditional independence. With high-dimensional data, sparsity of the precision matrix is often assumed, and various regularization methods have been applied for estimation. Under quite a few important scenarios, it is desirable to conduct the joint estimation of multiple precision matrices. In joint estimation, entries corresponding to the same element of multiple precision matrices form a group, and group regularization methods have been applied for estimation and identification of the sparsity structures. For many practical examples, it can be difficult to interpret the results when parameters within the same group have conflicting signs. To tackle this problem, we develop a regularization method for the joint estimation of multiple precision matrices. It effectively promotes the sign consistency of group parameters and hence can lead to more interpretable results, while still allowing for conflicting signs to achieve full flexibility. Its consistency properties are rigorously established. Simulation shows that the proposed method outperforms the competing alternatives under a variety of settings. With two data example, the proposed method leads to different and more consistent findings.

E0384: **Meta-clustering with multi-level omics data for cancer subtype discovery**
Presenter: **Yingying Wei**, The Chinese University of Hong Kong, Hong Kong
In traditional meta-analysis, we pool effect sizes across studies to improve statistical power. In meta-clustering, we want to conduct clustering jointly across studies. We propose a Bayesian hierarchical model that integrates diverse data types for clustering, accounts for the technical artifacts in individual studies, and handles cluster imbalance across studies. We apply the proposed methods to TCGA data and systematically identify subtypes for major human cancers.

EO103 Room LT-13 FINANCIAL STATISTICS Chair: Sheng-Feng Luo

E0272: **Systemic risk and interbank lending**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan
A simple model is proposed for the banking system incorporating a game feature where the evolution of monetary reserve is modeled as a system of coupled Feller diffusions. The Markov Nash equilibrium generated through minimizing the linear quadratic cost subject to Cox-Ingersoll-Ross type processes creates liquidity and deposit rate. The adding liquidity leads to a flocking effect but the deposit rate diminishes the growth rate of the total monetary reserve causing a large number of bank defaults. In addition, the corresponding Mean Field Game and the infinite time horizon stochastic game with the discount factor are also discussed.

E0382: **Marketability and discrete options with jump risk**
Presenter: **Sheng-Feng Luo**, Chung Yuan Christian University, Taiwan
Co-authors: Cheng-Der Fuh, Steven Kou, Hsinchieh Wong
A simple model based on lookback options by Longstaff has been widely used to study the value of marketability of a security, and has good empirical supports. However, a puzzle is why the model works so well, even if it ignores many practical features, such as discrete monitoring of the lookback options and jump risk. We find that although the discrete monitoring feature and the jump risk each has significant impacts on the model, interestingly the two effects tend to cancel each other, leading to the superb performance of the simple model. To reach this conclusion, we provide a general framework of approximating discrete monitoring options with jump risk, by significantly extending the Keener's method from diffusion models to jump diffusion models.

E0605: **The impact of fund characteristics and news sentiments on attention-flow relation**
Presenter: **Li-Jiun Chen**, Feng Chia University, Taiwan
Co-authors: Tsung-Ju Lee
The impact of investor attention on financial market has attracted many researchers attention. A general finding shows that mutual funds with

high attention are rewarded with positive flow of capital into the fund in subsequent periods. However, the issue of whether the sensitivity of this flow differs across fund characteristics and news sentiments remains open questions. We seek to address this issue by investigating and modeling fund characteristics and news sentiments affecting the relation between investor attention and capital flow. Furthermore, we explore whether this attention-flow relation is value-creating for investors. We empirically test the impact of fund characteristics and news sentiments on attention-flow relation and the impact of attention-flow relation behavior on investors wealth with a survivor-bias-free dataset that is available at the monthly horizon.

E0590: **Long-term and short-term impacts of common factors on correlated defaults**
Presenter: **Chu-Lan Kao**, National Chiao-Tung University, Taiwan
Co-authors: Charles Chang, Cheng-Der Fuh
Empirical studies have shown that the default of a firm impacts other firms in various ways. While some defaults only trigger short-term shock across the market, others tend to have a long-term effect on the correlated default structure. We utilize these two different features under the commonly used one-factor structural form model. In particular, through renewal theory, we show that a common factor without latent variable would only create short-term co-movement effect, but one with a latent variable will create a long-term impact. The different structure of the common factor is shown to make an essential difference on the correlate default dynamic.

EO174 Room LT-14 ROBUST LEARNING IN HIGH DIMENSIONAL DATA Chair: Guodong Li

E0284: **Quantile LASSO with changepoints in panel data models**
Presenter: **Matus Maciak**, Charles University, Czech Republic
Panel data models are quite modern statistical tools and they are commonly used in all kinds of econometric problems. In our approach we consider panel data models with changepoints, and atomic pursuit methods are utilized to detect and estimate these changepoints in the model. In order to obtain robust estimates and, also, to have a more complex insight into the data, we adopt the quantile lasso approach and the final model is obtained in a fully data-driven manner in just one modelling step. The main theoretical results are presented and some inferencial tools for changepoint significance are proposed. The presented methodology is applied for a real data scenario and some finite sample properties are investigated via s simulation study.

E0333: **Estimation of a two-component semiparametric location-shifted mixture model**
Presenter: **Jingjing Wu**, University of Calgary, Canada
Co-authors: Weixin Yao, Sijia Xiang, Xiaofan Zhou
Two efficient and robust estimators are discussed for a two-component semiparametric mixture model where the two components are unknown location-shifted symmetric distributions. Our estimators are derived by minimizing either the Hellinger distance (MHD) or the profile Hellinger distance (MPHD) between the model and a nonparametric density estimation. We propose simple and efficient algorithms to find the proposed MHD and MPHD estimators. Simulation studies are conducted to examine the finite sample performance of the proposed estimators and procedures and to compare them with other existing methods. We observe from our empirical studies that the two proposed estimators work very competitively with the existing methods for normal mixtures and much better for non-normal mixtures. More importantly, the proposed estimators are robust when data are contaminated with outlying observations. A real data is analyzed to illustrate the application of the proposed estimators.

E0365: **A robust and efficient approach to causal inference based on sparse sufficient dimension reduction**
Presenter: **Shujie Ma**, University of California-Riverside, United States
Estimation of treatment effects with a large number of covariates has received considerable attention in recent years. Most of the existing methods require specifying certain parametric models involving the outcome, treatment and confounding variables, and employ a variable selection procedure to identify confounders. However, selection of the right set of confounders depends on correct specification of the working models. The bias due to model misspecification and incorrect selection of confounders can yield misleading results. We proposes a new robust and efficient approach for inference about the average treatment effect via a flexible modeling strategy incorporating penalized variable selection. Specifically, we consider an estimator constructed based on an efficient influence function which involves a propensity score function and an outcome regression function. We then propose a new sparse sufficient dimension reduction approach to estimating these two functions, without making restrictive parametric modeling assumptions. We show that the proposed estimator of the average treatment effect is asymptotically normal and semiparametric efficient.

E0434: **Support recovery for sparse high dimensional matrices**
Presenter: **Adam Kashlak**, University of Alberta, Canada
Co-authors: Linglong Kong
The estimation problem for high dimensional covariance matrices and coefficient matrices is considered under the assumption of sparsity, which is qualitatively when most of the matrix entries are zero or negligible. Much past work has gone into such estimation based on approaches such as thresholding and LASSO. Instead, we take a unique and nonasymptotic approach to such estimation by using concentration inequalities to construct confidence sets for matrix estimators guaranteed to hold for finite samples. This is followed by an optimization over the confidence set in order to improve the estimator with respect to the sparsity criterion. In the context of support recovery, this methodology allows for the fixing of a false positive rate–i.e. zero entry claimed to be non-zero–and optimizing the true positive rate.

EO210 Room LT-15 BAYESIAN INFERENCE FOR STOCHASTIC FRONTIER MODELS Chair: Xibin Zhang

E0372: **Shadow prices of CO2 emissions: A random-coefficient, random-directional-vector directional distance function approach**
Presenter: **Guohua Feng**, University of North Texas, United States
The aim is to estimate the shadow prices of CO2 emissions of electric utilities in the US over the period from 2001 to 2014, using a random-coefficient, random-directional vector directional output distance function (DODF) model. The main feature of this model is that both its coefficients and directional vector are allowed to vary across firms, thus allowing different firms to have different production technologies and to follow different growth paths. Our Bayes factor analysis indicates that this model is strongly favored over the commonly used fixed-coefficient DODF model. Our results obtained from this model suggest that the average annual shadow price of CO2 emissions ranges from \$61.62 to \$105.72 (in 2001 dollars) with an average of \$83.12. The results also suggest that the firm-specific average shadow price differs significantly across electric utilities. In addition, our estimates of the shadow price of CO2 emissions show an upward trend for both the sample electric utilities as a whole and the majority of the individual sample electric utilities.

E0452: **Bayesian stochastic frontiers using transformation to normal**
Presenter: **Reza Hajargasht**, Swinbuurne University of Technology, Australia
The normal distribution is very convenient. In particular, one can transparently allow for correlation between normally distributed variables and also conditional distributions can be easily defined for multivariate normal distributions. One cannot, however, directly specify a normal distribution

for an efficiency effect in stochastic frontier models due to the fact the distribution of the effects must be one-sided. But, it is always possible to transform a one-sided random variable with a known distribution to another that is normally distributed. The purpose is to show how using these simple facts (i.e. by transforming the efficiency effects "ui" to a normally distributed variable and allowing for correlation between transformed variables), one is able to handle some difficult problems in stochastic frontier analysis relatively easily. We consider problems such as stochastic frontiers with endogeneity, Stochastic frontiers with serially correlated errors and Stochastic frontier models with factor error structure and show how they can be estimated using either Bayesian or maximum simulated likelihood approaches.

E0546: **Bayesian estimation of dynamic stochastic frontier model: A simulation study**
Presenter: **Chuan Wang**, Zhongnan University of Economics and Law, China
A stochastic frontier model is proposed that allows for long memory dynamic technical inefficiency structure. We use AR(p) model to explore this temporal behaviour of inefficiency in a panel data setting. We also propose a MCMC method to estimate the lag in the model. To compare the performance of our method with the mainstream Bayesian lag selection method (i.e. Bayes factor), a comprehensive simulation study is conducted.

E0654: **Panel data analysis of hospital variations in length of stay for hip replacements: Private versus public**
Presenter: **Xibin Zhang**, Monash University, Australia
Co-authors: Yan Meng, Xueyan Zhao, Jiti Gao
Inequality between private and public patients in Australia has been an ongoing concern due to its two tiered insurance system. The aim is to investigate the variations in hospital length of stay for hip replacements using the Victorian Admitted Episodes Dataset from 2003-2004 to 2014-2015, employing a Bayesian hierarchical random coefficient model with trend allowing for structural break. We find systematic differences in the length of stay between public and private hospitals, after observable patient complexity is controlled. This suggests shorter stays in public hospitals due to pressure from the Activity-based funding scheme, and longer stays in private system due to potential moral hazard. Our counterfactual analysis shows that public patients stay 1.4 days shorter than private in 2014, which leads to the quicker but sicker concern that is commonly voiced by the public. We also identify widespread variations among individual hospitals. Sources for such variation warrant closer investigation by policy makers.

EO020 Room LT-16 RECENT ADVANCES IN COMPLEX DATA ANALYSIS Chair: Xingqiu Zhao

E0236: **Two modeling strategies for two-part latent variable model**
Presenter: **Yemao Xia**, Nanjing Forestry University, China
Semi-continuous data often occur in the survey of economics and social sciences. In analyzing such data, a primary interest is to assess the the effects of observed covariates on the variability of responses. We extend the two-part regression model to the case where the unobserved heterogeneities are explained by the latent variable model. The information on latent factors is specified via latent variable model. We develop two estimation procedures for analyzing such data: one is based on robust moment estimation equation and the other is within the Bayesian framework. For the former, we establish two-step estimation procedure for the unknown parameters involved and investigate the asymptotic properties such as consistency and asymptotic normality; while for the latter, we design a Polya-Gamma Gibbs sampler in the Bayesian posterior sampling. We also assess model fits via constructing various related hypothesis testing procedures. Simulation studies were carried out to assess the performance of the two approaches, especially the robust behavior of estimates and tests when the underlying distribution assumptions are violated. A real data set is analyzed to illustrate the practical values of the proposed methodology.

E0269: **Bayesian semiparametric quantile regression modeling for estimating earthquake fatality risk**
Presenter: **Yunxian Li**, Yunnan University of Finance and Economics, China
Earthquake often results in significant life and property losses. Due to its limitation in analyzing catastrophic loss, mean regression may not be appropriate for analyzing fatality risk caused by earthquake. We developed a Bayesian semiparametric quantile regression model for count data. The count responses are converted to continuous responses through the jittered method and a transform function. A Bayesian semiparametric quantile regression modeling approach is then developed. The error distribution in the quantile regression model is assumed to be a mixture of asymmetric Laplace distributions constructed with Dirichlet process. Historical death tolls of China caused by earthquakes from 1969 to 2006 are used for fitting and a parametric model is employed for model comparison. The results of model comparison show that the proposed semiparametric quantile regression model outperforms the parametric model. The empirical analysis illustrates that the impact of earthquake magnitude on death tolls is significant. Moreover, the impact of the magnitude is more pronounced on higher percentiles of death tolls.

E0271: **A new regression method**
Presenter: **Pengfei Liu**, Jiangsu Normal University, China
Mean regression and quantile regression are popular statistical method. We consider the distance of characteristic functions, then propose a new method to estimate parameters of linear regression. We also discuss the variable selection methods. The proposed method are illustrated by some simulated data and real data.

E0287: **Zero-inflated regime-switching stochastic differential equation models for multivariate multi-subject time-series data**
Presenter: **Zhaohua Lu**, St. Jude Children's Research Hospital, United States
Co-authors: Sy-Miin Chow, Nilam Ram, Pamela Cole
Stochastic differential equation (SDE) models are widely used in the studies of human dynamics, which are often characterized by the sparse occurrences of certain behavior in some individuals. To recover the dynamics of a system with an inflation of such zero responses, we incorporate a regime (latent phase) of non-occurrence to an SDE model to account for the high proportion of non-occurrence instances and simultaneously model the multivariate dynamic processes of interest under non-zero responses. The transition between the occurrence and non-occurrence regimes is represented by a latent Markovian transition model which depends on latent regime indicators and person-specific covariates. Markov chain Monte Carlo algorithms are used for the Bayesian estimation and inference. We demonstrate the proposed zero-inflated regime-switching SDE model through a multi-subject dynamic self-regulation study for young children at 36 and 48 months.

EO095 Room LT-17 ADVANCES IN NONLINEAR AND FINANCIAL TIME SERIES Chair: Wai-Keung Li

E0281: **Buffered vector error-correction models**
Presenter: **Philip Yu**, The University of Hong Kong, Hong Kong
Co-authors: Renjie Lu

The buffered autoregressive model is extended to the buffered vector error correction model (VECM). Least squares estimation and a reduced-rank estimation are discussed, and the consistency of the estimators on the delay parameter and threshold parameters is derived. We also propose a supWald test for the presence of buffer-type threshold effect. Under the null hypothesis of no threshold, the supWald test statistic converges to a function of Gaussian process. A bootstrap method is proposed to obtain the p-value for the supWald test. We investigate the effectiveness of our methods by simulation studies. We apply our model to study the monthly Federal bond rates of United States and identify evidences of buffering regimes in the bond rates.

E0207: **A novel partial-linear single-index model for time series data**
Presenter: **Hui Jiang**, Huazhong University of Science and Technology, China
Co-authors: Lei Huang

Partial-linear single-index models have been widely studied and applied, but its current applications to time series modelling still need some strong and inappropriate assumptions. A novel method is proposed which relaxes those assumptions and extend the applicability of partial-linear single-index models to time series modeling, taking both lag variables and autocorrelated errors into consideration. An estimation procedure based on Whittle likelihood is proposed and some asymptotic properties of the estimators are derived. We conduct some simulations to elaborate that the proposed model and methodology is necessary in certain situations. These situations are shown to be common and rational by two examples of real data analyses, which also consequently show the feasibility and practicability of our proposed model and approach.

E0475: **On Brownian motion approximation of compound Poisson processes with applications to threshold models**
Presenter: **Dong Li**, Tsinghua University, China

Compound Poisson processes (CPP) constitute a fundamental class of stochastic processes and a basic building block for more complex jump-diffusion processes such as the Levy processes. However, unlike those of a Brownian motion (BM), distributions of functionals, e.g. maxima, passage time, argmin and others, of a CPP are often intractable. We propose a novel approximation of a CPP by a BM so as to facilitate closed-form expressions in concrete cases. Specifically, we approximate, in some sense, a sequence of two-sided CPPs by a two-sided BM with drift, when the threshold effects (e.g. the differences between the regression slopes in two-regime regression) are small but fixed. As applications, we use our approximation to perform statistical inference of threshold models, such as the construction of confidence intervals of threshold parameters. These models include threshold regression (also called two-phase regression or segmentation) and numerous threshold time series models. We conduct numerical simulations to assess the performance of the proposed approximation. We illustrate the use of our approach with a real data set.

E0448: **Random weighting the Portmanteau tests for multivariate white noise with unknown dependent structure**
Presenter: **Yanfen Zhang**, Xiamen University, China

The random weighting method is extended from univariate to multivariate to bootstrap the critical value of Ljung-Box portmanteau test in multivariate weakly white noise. A set of Monte Carlo experiments that check multivariate white noise illustrate the practical relevance of this method. A real example on the USD-MYR and USD-SGD five day rate of return on foreign exchange illustrates the merits of our testing procedure.

EO127 Room LT-18 RECENT ADVANCES IN TIME SERIES ANALYSIS Chair: Chi Tim Ng

E0292: **New active zero set descent algorithm for LAD problems with generalized lasso penalty**
Presenter: **Yue Shi**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Zhiguo Feng, Chi Tim Ng, Cedric Yiu

A new active zero set descent algorithm is proposed for least absolute deviance(LAD) problems with generalized Lasso penalty. Zero set contains the terms in the cost function that are zero-valued at the solution. Unlike state-of-art numerical approximation strategies such as interior point method, user-chosen threshold value is not required by the proposed algorithm to identify the zero set. Moreover, no nested iteration is needed. The algorithm updates the zero set and basis search directions recursively until optimality conditions are satisfied. It is also shown that the proposed algorithm converges in finitely many steps. Extensive simulation studies and real data analysis are conducted to confirm the time-efficiency of our algorithm.

E0419: **Modeling and forecasting online auction price**
Presenter: **Weiwei Liu**, Lanzhou university, China

Online auctions have become increasingly popular in recent years. There is a growing body of research on this topic, whereas modeling online auction price curve constitutes one of the most interesting problems. Most of research treats price curves as deterministic functions, which ignores the random effects of external and internal factors. To account for the randomness, a more realistic model using stochastic differential equation is proposed. The online auction price is modeled by a stochastic differential equation in which the deterministic part is equivalent to a previous second-order differential equation model proposed. The model also includes a component representing the measurement errors. Explicit expressions for the likelihood function are also obtained, from which statistical inference can be conducted.

E0416: **Exterior point algorithm for change point analysis of general time series models**
Presenter: **Chi Tim Ng**, Chonnam National University, Korea, South

An efficient exterior point algorithm is proposed for smoothing and change point detection of financial time series data under the penalized likelihood approach. The proposed method has $O(n)$ computational complexity and is applicable to a broad class of time series model proposed that encompasses ARMA-GARCH as a special case. Under certain conditions, the estimated model has piecewise constant coefficients. Asymptotic properties of the penalized likelihood estimators are established. The possibility of real-time forecasting that update the prediction within $O(1)$ time upon arrival of new signal is also discussed.

E0418: **MOSUM based test for variance change in panel data**
Presenter: **Man Wang**, Donghua University, China

The problem of testing for common variance changes in panel data model is considered. We propose a test based on moving sum (MOSUM) of squares. Asymptotic distribution of the test statistic is derived under the null hypothesis of no change and the consistency of the test is also established. We conduct several simulation studies, which show that the proposed test performs better than some existing CUSUM based tests when the magnitude of variance change is small. Additionally, in the presence of multiple changes, the variances of the panels may increase at one change point and decrease at another. The increase and decrease in variances may cancel each other out and result in loss of power of the CUSUM based test method. Under this circumstance, the new MOSUM based test has obvious superiority over these CUSUM based test, as shown in the simulation.

EO133 Room P4302 NETWORK ANALYSIS Chair: Frederick Kin Hing Phoa

E0265: **Efficient spread of networks**
Presenter: **Yuan-Lung Lin**, Academia Sinica, Taiwan
Co-authors: Frederick Kin Hing Phoa
The growth of social networks, in combination with the increasing sophistication of Big Data tools, has led to a burgeoning interest in a rich understanding of relationships among people, institutions, and more. A relevant setting for such a study is graph theory, together with its random counterpart. Many graph models have been employed to investigate the clusters of nodes and the centrality of each cluster based on structure and attributes. The centrality of a network is one of the key measure of the importance of the nodes with respect to the rest of the nodes, such as degree, betweenness, eigenvector, and closeness. Each centrality is used for different purposes, but none of them is applicable to spread information in a network. Our interest is in spreading information efficiently in a network. We will propose a new measurement, domination centrality sets, which combines the advantages of known methodologies without their drawbacks. Besides, a new algorithm based on domination centrality sets will be proposed for the search of important nodes with effective spreading. For this purpose, we will also derive some theoretical methodologies which can help users to avoid exhaustive and time consuming computation.

E0381: **Designing experiments for general network structures**
Presenter: **Ming-Chung Chang**, Graduate Institute of Statistics, National Central University, Taiwan
Co-authors: Frederick Kin Hing Phoa, Jing-Wen Huang
Experiments on connected units are commonly conducted in various fields, such as agriculture trials, medical experiments and social networks. In these cases, an experimental unit may connect with some others, and the treatment applied to a unit has an effect, called a network effect, on the responses of the neighboring units. Designing such experiments is rarely discussed in the literature. A study of A-optimal designs on connected experimental units with unstructured treatments has been previously initiated. It was assumed that the network effects are unknown constants. We study a similar design problem but assuming that those effects are random effects, which lead to a property that the responses of two units are correlated if some neighbors of one unit and those of the other receive the same treatment. Alphabetical optimality criteria are considered for selecting good designs with high efficiency of estimating the treatment effects and/or high accuracy of predicting the network effects. We provide theoretical conditions for designs to be optimal and illustrate our theory with some numerical examples.

E0488: **Estimating links of a network from time to event data**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan
A statistical method is developed for identifying links of a network from time to event data. This method models the hazard function of a node conditional on event time of other nodes, parameterizing the conditional hazard function with the links of the network. It then estimates the hazard function by maximizing a pseudo partial likelihood function with parameters subject to a user-specified penalty function and additional constraints. To make such estimation robust, it adopts a pre-specified risk control on the number of false discovered links by using the Stability Selection method. Simulation study shows that under this hybrid procedure, the number of false discovered links is tightly controlled while the true links are well recovered. We apply our method to estimate a political cohesion network that drives donation behavior of 146 firms from the data collected during the 2008 Taiwanese legislative election. The results show that firms affiliated with elite organizations or firms of monopoly are more likely to diffuse donation behavior. In contrast, firms belonging to technology industry are more likely to act independently on donation.

E0583: **A comparative study of academic papers on the PM2.5 environmental issues in China and Japan**
Presenter: **Yuji Mizukami**, Nihon University, Japan
Co-authors: Takao Nagai, Shigo Chin, Frederick Kin Hing Phoa, Keisuke Honda, Junji Nakano
Since the 1990s, China has been remarkably growing. On the other hand, environmental problems are getting worse, especially the damage of PM 2.5 is serious and early solution is required. Research on PM2.5 is remarkably active in China and many papers have been produced. Japan is working on pollution problems since the beginning of 1900 and has accumulated knowledge on pollution problems. As for the problem of PM2.5, it is studied as a foreign environmental problem. We will explore the characteristics of the PM 2.5 research in China and Japan by network analysis and consider the field of environmental research.

EO188 Room P4703 DISCRETE DATA ANALYSIS: PROBLEMS, CHALLENGES, AND SOLUTIONS Chair: Heping Zhang

E0459: **Bayesian modeling of multivariate non Gaussian time series**
Presenter: **Refik Soyer**, George Washington University, United States
Modeling of multivariate non Gaussian time series of correlated observations is considered. In so doing, we focus on time series from multivariate counts and durations. Dependence among series arises as a result of sharing a common dynamic environment. We discuss characteristics of the resulting multivariate time series models and develop Bayesian inference for them using particle filtering and Markov chain Monte Carlo methods.

E0342: **Residuals and diagnostics for ordinal regression models: A surrogate approach**
Presenter: **Dungang Liu**, University of Cincinnati, United States
Co-authors: Heping Zhang
Ordinal outcomes are common in scientific research and everyday practice, and regression models are often used to make inference. A long-standing problem with such regression analyses is the lack of effective diagnostic tools for validating model assumptions. The difficulty arises from the fact that an ordinal variable has discrete values that are labeled with, but not, numerical values. The values merely represent ordered categories. We propose a surrogate approach to defining residuals for an ordinal outcome Y. The idea is to define a continuous variable S as a "surrogate" of Y and then obtain residuals based on S. For the general class of cumulative link regression models, we study the residual's theoretical and graphical properties. We show that the residual has null properties similar to those of the common residuals for continuous outcomes. The numerical studies demonstrate that the residual has the power to detect misspecification with respect to 1) mean structures; 2) link functions; 3) heteroscedasticity; 4) proportionality; and 5) mixed populations. The results suggest that compared to a previously defined residual, our residual can reveal deeper insights into model diagnostics. We stress that the focus is on residual analysis, rather than hypothesis testing. The latter has limited utility as it only provides a single p-value, whereas our residual can reveal what components of the model are misspecified and advise how to make improvements.

E0300: **Partial association between ordinal variables: Quantification, visualization and estimation**
Presenter: **Shaobo Li**, University of Cincinnati, United States
Co-authors: Dungang Liu, Yan Yu
Partial association measures the relationship between two variables Y_1 and Y_2 after adjusting a set of covariates X. It remains unknown how to fully characterize such an association if both Y_1 and Y_2 are recorded on ordinal scales. We propose a general measure, labeled as ϕ, to characterize ordinal-ordinal partial association. It is based on surrogate residuals derived from fitting cumulative link regression models for each Y_1 and Y_2. We show the measure has the following properties: (1) its size reflects the strength of association for ordinal data, rather than the hypothetical

latent variables; (2) it does not rely on the normality assumption or models with the probit link, but instead it broadly applies to models with any link functions; and (3) it can capture non-linear association and has a potential to detect dependence of any complex structures. We stress that the focus is not on hypothesis testing, but quantification and visualization. We demonstrate that our numerical and graphic assessment can reveal microstructure of partial association, which can inform statistical modeling of multivariate ordinal data.

E0454: **Flexible joint models for correlated jittered discrete data via Gaussian copulas**
Presenter: **Alexander de Leon**, University of Calgary, Canada
Although a latent variable description of discrete variables is practically appealing in many applications, such an approach does not work in all cases. Count data, for example, present a situation for which a latent variable framework may not be appropriate; this is also the case for nominally scaled categorical outcomes, such as gender and hair colour, for example. We adopt the jittering method (or "continuous-ation") in order to circumvent the complications engendered by the direct application of copulas to discrete variables (e.g., non-identifiability, non-margin-free dependence). The method entails transforming discrete data into continuous, hence the name, by the addition of independently generated continuous random variables, called "jitters", for which a joint model is constructed using the Gaussian copula. Such an approach to joint modelling of correlated discrete data is appealing in practice because it preserves the dependence structure of the data, in that the associations among discrete variables are the same as those between their corresponding jittered versions. We discuss likelihood estimation for such models and report simulation results on the finite-sample relative bias and efficiency of resulting estimates. We adopt the approach to develop spatial joint models for areal count data on the numbers of deaths due to cancers of lung and esophagus during the period 1991-1998 in the 87 counties of Minnesota, USA.

EO123 Room P4704 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E0216: **Optimal design for mixed effects models**
Presenter: **John Stufken**, Arizona State University, United States
Identifying optimal designs for correlated data is a difficult problem. Many classical results for independent data have no obvious generalization to correlated data. We propose a method to identify locally optimal designs for classes of linear, generalized linear, and nonlinear mixed effects models under commonly used optimality criteria by extending results for independent data. We demonstrate the method through a real life study, and investigate robustness of design efficiency to mis-specification of the covariance matrix for the random effects.

E0255: **Robust dose-level designs for binary responses in environmental risk assessment**
Presenter: **Wanchunzi Yu**, Bridgewater State University, United States
Co-authors: John Stufken
The estimation of the lower confidence limit of a benchmark dose (BMD) is a common objective in environmental risk assessment. A risk function is involved to determine the BMD that corresponds to a given, low-level benchmark response (BMR). In a BMD study, measurements are taken at different dose levels for a pollutant of interest. These dose levels need to be selected in a controlled experiment, which is a design problem. We proposed a weighted c-efficiency criterion to obtain a design which is robust to risk function uncertainty and misspecification. Furthermore, a min-max weighted c-efficiency criterion is also developed to find a design not only robust to various risk functions, but also to different function parameters. In the simulation studies, we apply the particle swarm optimization (PSO) algorithm to search for designs. The presented methods for identifying robust designs is also demonstrated through the mammalian carcinogenicity of cumene (C9H12) example.

E0295: **Optimal design of sampling survey for efficient parameter estimation**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: Xueqin Wang
For many tasks of data analysis, only the information of the explanatory variable may be available and the evaluation of the response values are quite expensive. While it is impractical or too costly to obtain the responses of all units, a natural remedy is to judiciously select a good sample of units, for which the responses are to be evaluated. We adopt the classical criteria in design of experiments to quantify the information of a given sample regarding parameter estimation. Then, we provide a theoretical justification for approximating the optimal sample problem by a continuous problem, for which fast algorithms can be further developed with the guarantee of global convergence. Our results have the following novelties: (i) The statistical efficiency of any candidate sample can be evaluated without knowing the exact optimal sample; (ii) It can be applied to a very wide class of statistical models; (iii) It can be integrated with a broad class of information criteria; (iv) It is much faster than existing algorithms. (v) A geometric interpretation is adopted to theoretically justify the relaxation of the original combinatorial problem to continuous optimization problem.

E0504: **Optimal experimental designs for ultra-fast brain imaging studies**
Presenter: **MingHung Kao**, Arizona State University, United States
Neuroimaging technologies such as functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS) have been widely used in various fields, including economics, neuroscience, political science and psychology, for studying the inner working of the human brain. Recent advances in these technologies allow researchers to collect neuroimaging data with improved temporal resolutions. However, obtaining optimal designs for such ultra-fast imaging studies is computationally very difficult, if not infeasible. We propose a computational method that makes it possible to efficiently obtain optimal experimental designs for collecting highly informative data from ultra-fast neuroimaging studies.

Thursday 21.06.2018 16:00 - 17:15 Parallel Session M – EcoSta2018

EO263 Room B4302 HIGH-DIMENSIONAL ESTIMATION IN ECONOMETRICS Chair: Zhentao Shi

E0224: **An exact and robust conformal inference method for counterfactual and synthetic controls**
Presenter: **Yinchu Zhu**, University of Oregon, United States
Co-authors: Victor Chernozhukov, Kaspar Wuthrich
New inference methods are introduced for counterfactual and synthetic control methods for evaluating policy effects. Our inference methods work in conjunction with many modern and classical methods for estimating the counterfactual mean outcome in the absence of a policy intervention. Specifically, our methods work together with the difference-in-difference, canonical synthetic control, constrained and penalized regression methods for synthetic control, factor/matrix completion models for panel data, interactive fixed effects panel models, time series models, as well as fused time series panel data models. The proposed method has a double justification. (i) If the residuals from estimating the counterfactuals are exchangeable as implied, for example, by i.i.d. data, our procedure achieves exact finite sample size control without any assumption on the specific approach used to estimate the counterfactuals. (ii) If the data exhibit dynamics and serial dependence, our inference procedure achieves approximate uniform size control under weak and easy-to-verify conditions on the method used to estimate the counterfactual. We verify these condition for representative methods from each group listed above. Simulation experiments demonstrate the usefulness of our approach in finite samples. We apply our method to re-evaluate the causal effect of election day registration (EDR) laws on voter turnout in the United States.

E0246: **Boosted panel data approach for program evaluation**
Presenter: **Zhentao Shi**, CUHK, Hong Kong
Policy evaluation is a central question in empirical economic studies, but economists mostly work with observational data in view of the limited opportunities to carry out controlled experiments. The lack of genuine control groups motivated to exploit the correlation between cross-sectional units in a panel data to construct the counterfactual. The choice of cross-sectional units, a key step in implementing such a method, has not been not addressed in the case of many potential controls. We propose to use the component-wise boosting for control-unit selection. We show that such a choice is asymptotically valid even if the number of potential controls grows, in the limit, faster than the time dimension. Both in theory and in practice, we open the possibility the above-mentioned method to be applied to empirical research in big data environment.

E0301: **Forecast combinations for predictive regressions via the Lasso**
Presenter: **Bonsoo Koo**, Monash University, Australia
Co-authors: Hong Wang
When a number of specifications are suggested, forecast combination reduces the information in a vector of forecasts to a single summary measure using a set of combination weights. While the reasons are poorly understood, simple equal weighted (EW) forecast combination scheme often outperforms more sophisticated combination schemes in empirical studies. We propose a Least Absolute Shrinkage and Selection Operator (LASSO) estimator of the optimal combination weights of which are estimated from potentially highly correlated covariates (individual forecasts). Motivated by the properties of LASSO, we demonstrate two applications of the proposed LASSO approach in time series setting. The proposed LASSO approach is applied to forecasting stock returns with comparison to the simple equal weighted (EW) combination scheme, which in turn outperforms the best individual predictions.

EO176 Room G4302 MODEL AVERAGING Chair: Tian Xie

E0296: **Model averaging estimation for conditional heteroscedasticity model family**
Presenter: **Qingfeng Liu**, Otaru University of Commerce, Japan
Co-authors: Qingsong Yao, guoqing Zhao
The model averaging estimation for the conditional heteroscedasticity model family is considered. Given a set of candidate models with different functional forms, we propose a model averaging estimator for the conditional volatility and construct the corresponding weight choosing criterion. According to our results, the weight that minimizes the weight choosing criterion asymptotically minimizes the true KL divergence, as well as the Itakura-Saito distance.

E0167: **Twits versus tweets: Does adding social media wisdom Trump admitting ignorance when forecasting the CBOE VIX?**
Presenter: **Tian Xie**, WISE, Xiamen University, China
Co-authors: Steven Lehrer, Xinyu Zhang
A rapidly growing literature has documented improvements in forecasting financial return volatility measurement via use of variants of the heterogeneous autoregression (HAR) model. At the same time, there is an increasing number of products made from social media that are suggested to improve forecast accuracy. We first develop a model averaging heterogeneous autoregression (MAHAR) model that can account for model uncertainty. Second, we use a deep learning algorithm on a 10% random sample of Twitter messages at the hourly level to construct a sentiment measure that is being marketed by the Wall Street Journal. Our empirical results suggest that jointly incorporating model averaging techniques and sentiment measures from social media can significantly improve the forecasting accuracy of financial return volatility.

E0311: **Time-varying model averaging**
Presenter: **Yuying Sun**, Academy of Mathematics and System Science, Chinese Academy of Sciences, China
Co-authors: Xinyu Zhang, Tae-Hwy Lee, Yongmiao Hong, Shouyang Wang
Structural changes often occur in economics and finance due to changes in preferences, technologies, institutional arrangements, policies, crises, etc. Improving the forecast accuracy of economic time series with the evolutionary behavior is a long-standing problem. Model averaging aims at providing an insurance against selecting a poor model. All existing model averaging approaches are designed with constant weights. Little attention has been paid to the time-varying model averaging, which is more realistic in economics under structural changes. A novel model averaging estimator is proposed which selects the smoothly time-varying weights by minimizing a local jackknife criterion. It is shown that the proposed time-varying jackknife model averaging (TJMA) estimator is asymptotically optimal in the sense of achieving the lowest possible local squared errors in a class of time-varying model averaging estimators, with allowing non-spherical errors. A simulation study and empirical application highlight the merits of the proposed TJMA estimator relative to a variety of popular estimators from constant model averaging and model selection.

EO267 Room G4701 NEW DEVELOPMENT IN FUNCTIONAL DATA ANALYSIS Chair: Tiejun Tong

E0276: **Semi-parametric modeling of structured point processes using multi-level log-Gaussian Cox processes**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States
Co-authors: Ming Wang, Ganggang Xu, Hui Huang, Yongtao Guan

A general framework is proposed for using multi-level log-Gaussian Cox processes to model repeatedly observed point processes with complex structures. A novel nonparametric approach is developed to consistently estimate the covariance kernels of the latent Gaussian processes at all levels. Consequently, multi-level functional principal component analysis can be conducted to investigate the various sources of variations in the observed point patterns. In particular, to predict the functional principal component scores, we propose a consistent estimation procedure by maximizing the conditional likelihoods of super-positions of point processes. We further extend our procedure to the bivariate point process case where potential correlations between the processes can be assessed. Asymptotic properties of the proposed estimators are investigated, and the effectiveness of our procedures is illustrated by an application to a stock trading dataset.

E0508: **Functional censored quantile regression with application to the BOSS data analysis**
Presenter: **Fei Jiang**, The University of Hong Kong, Hong Kong

A functional censored quantile regression model is proposed to study the relationship between the time to stroke recurrence and the dynamic blood pressure levels, where the time-varying effect is an unspecified function. The B-spline method is used to approximate the coefficient function, which operationally reduces the problem to the parametric estimation. A generalized approximate cross-validation method is developed to select the number of knots by minimizing the expected loss. We demonstrate the asymptotic properties of the estimation and knots selection procedure. Further, we conduct extensive simulation to evaluate the finite sample performance of our method and apply it to analyze the blood pressure and clinical outcome in transient ischemic attack or ischemic stroke data. The results reinforce the importance of the morning surge phenomenon, whose effect has caught attention but remains controversial in the medical literature.

E0639: **Estimation and classification for varying-coefficient panel data model with latent structures**
Presenter: **Tao Huang**, Shanghai University of Finance and Economics, China

A varying coefficient panel data model with unknown group structures is considered, where the group membership of each individual and the number of groups are left unspecified. We first develop a triple localization approach to estimate the unknown coefficient functions, and then identify latent grouped structures via community detection method. To improve the efficiency of the resultant estimator, we further propose a two-stage estimation method that enables the resulting estimator achieve optimal rates of convergence. In the theoretical part, the asymptotic theory of the resultant estimators are derived. In particular, we provide the convergence rates and the asymptotic distribution of our estimators. In the empirical part, several simulated examples and a real data analysis are presented to illustrate the finite sample performance of the proposed methods.

EO269 Room LT-11 CORPORATE BOND LIQUIDITY AND CREDIT RISKS Chair: Yuan Wang

E0257: **Liquidity risk and corporate risk-taking**
Presenter: **Yuan Wang**, Concordia University, Canada
Co-authors: Jingzhi Huang, Rui Zhong

A theoretical framework is constructed to investigate the impact of liquidity risk, in the secondary corporate debt market, on corporate risk-taking preferences. Using closed-form solutions, our model shows that equity holders choose to adopt high-risk projects upon the arrival of illiquidity shocks. This effect is more pronounced for firms with weaker fundamentals. Empirically, we confirm the positive relationship between liquidity risk and corporate risk-taking. We also document that the impact of liquidity risk on corporate risk-taking preferences is more pronounced for smaller firms and firms with lower profits and higher rollover risk. In addition, we use the introduction of the Trade Reporting and Compliance Engine (TRACE) as a natural exogenous liquidity shock and find a decrease of firms risk-taking preferences after the TRACE is implemented. Our findings shed light on the managerial behavior literature, which shows that the frictions of the secondary bond market have a real impact on firms risk-taking behaviors.

E0334: **Effects of the single stock circuit breaker on the stock market: Canadian evidence**
Presenter: **Lorne Switzer**, Concordia University, Canada

The purpose is to look at the effects of regulatory changes in which restrictions on short sales on stocks with declining prices are replaced by circuit breakers that are triggered when individual stocks experience large upside or downside movements. The focus is on all stocks traded on the Toronto Stock Exchange since the inception of the single stock circuit breaker rule in February 2012. The results reject the delayed price discovery hypothesis since the material information that caused the circuit breaker induced trading halt is incorporated in stock prices at the time of the halt. We find that with the exception of implied volatility, intraday volatility measures decline for stocks affected by the circuit breaker. This volatility decline is not associated with reduced market liquidity, however.

E0774: **Bank activism and value creation**
Presenter: **Jun Wang**, University of Western Ontario, Canada
Co-authors: Kee-Hong Bae, Keke Song

The aim is to investigate the impact of bank activism on target firms debtholders and shareholders by examining the abnormal bond and stock returns around shareholder activism events. We find that debtholders rather than shareholders benefit when shareholder activists are banks. We also find that relative to other activists, bank activists are more likely to target larger financial firms with higher leverage and lower credit quality and that bank activism target firms experience greater reduction in leverage ratio and improvement in credit quality. Additionally, the positive abnormal bond returns associated with bank activism only exist in the subsample of bank activism events where bank shareholder activists are also current lenders of the same target firms. We interpret these findings as follows: bank activists gain control rights through delegation of their trust business clients; the separation of cash rights and control rights associated with banks proxy holdings may cause a conflict of interests problem if bank activists also hold loan stake in the same target firms.

EO160 Room LT-12 STATISTICAL INFERENCE IN HIGH DIMENSIONAL QUANTILE REGRESSION Chair: Yanlin Tang

E0184: **A conditional marginal test in high-dimensional quantile regression**
Presenter: **Yanlin Tang**, TONGJI University, China
Co-authors: Yinfeng Wang, Huixia Wang, Qing Pan

A conditional marginal score-type test in high-dimensional quantile regression is proposed in order to test the presence of significant covariates given a conditioning set. The test is based on the maximal score-type test statistics, and under mild regularity conditions, the proposed test statistic converges to a type I extreme value distribution, after some standardization. Besides the asymptotic distribution, we also propose a multiplier bootstrap method for critical value construction. We also illustrate how the proposed test can be used as a stopping rule in forward regression. We show, through simulation, that the proposed method provides adequate control of the family-wise error rate with competitive power. We illustrate the application of our method by analyzing a GFR data.

E0215: **Quantile-regression-based clustering for panel data**
Presenter: **Yingying Zhang**, Fudan University, China
Co-authors: Huixia Judy Wang, Zhongyi Zhu

In many applications, such as economic and medical studies, it is important to identify subgroups of subjects with different covariate effects. We propose a new quantile-regression-based clustering method for panel data. We develop an iterative algorithm using a similar idea of k-means clustering to identify subgroups at a single quantile level or at multiple quantiles jointly. Even in cases where the group membership is the same across quantile levels, the signal differentiating subgroups may vary with quantiles. It remains unclear which quantile is preferable or should one use composite regression by combining information across multiple quantiles. To answer this question, we propose a new stability measure to choose among multiple quantiles and the composite quantile that gives the most stable clustering results. The consistency of the proposed parameter and group membership estimation is established. The finite sample performance of the proposed method is assessed through simulation and the analysis of an economy growth data.

E0353: **Direction estimation in single-index quantile regressions via martingale difference divergence**
Presenter: **Jicai Liu**, Shanghai Normal University, China

A novel estimation method based on the martingale difference divergence in single index quantile models is proposed. Our approach does not require any nonparametric estimation and enjoys a model free property. Under regularity conditions, we show that our estimator is root-n consistent and asymptotically normal. We compare the performance of our method with the single index estimation method by simulations and show that our method is very competitive and robust across a number of models. Finally, we analyze a real data set to demonstrate the efficacy of our method.

EO026 Room LT-13 STATISTICAL LEARNING IN FINANCE Chair: Guanhao Feng

E0391: **Latent common return volatility factors: Capturing elusive predictive accuracy gains when forecasting volatility**
Presenter: **Mingmian Cheng**, Rutgers University, United States
Co-authors: Norman Swanson, Xiye Yang

Factor-augmented HAR-type models are used to predict the daily integrated volatility of asset returns. Our approach is based on a proposed two-step dimension reduction procedure designed to extract latent common volatility factors from a large dimensional and high-frequency return data set with 267 constituents of the S&P 500 index. In the first step, we apply either Lasso or elastic net shrinkage on estimates of integrated volatility of all constituents in the data set, in order to select a subset of asset return series for further processing. In the second step, we utilize (sparse) principal component analysis to estimate latent common asset return factors, from which latent integrated volatility factors are extracted. Although we find limited in-sample fit improvement, relative to a benchmark HAR model, all of our proposed factor-augmented models result in substantial out-of-sample predictive accuracy improvement. In particular, forecasting gains are observed at market, sector, and individual-stock levels, with the exception of the financial sector. Further investigation of the factor structures for non-financial assets shows that industrial and technology stocks are characterized by minimal exposure to financial assets, inasmuch as forecasting gains associated with factor-augmented models for these types of assets are largely attributable to the inclusion of non-financial stock price return volatility in our latent factors.

E0665: **Term structure of recession probabilities and the cross section of asset returns**
Presenter: **Ti Zhou**, Southern University of Science and Technology, China

The duration of business cycles changes over time, generating time-varying investor concern about recessions. We study a new macro-factor model that directly links assets' risk premia to such a concern, measured by the term structure of recession probabilities from professional forecasters. The innovation to the slope of the term structure is negatively priced with an economically large and significant risk premium in a wide range of tests assets, consistent with how the slope predicts long-run macroeconomic activity and labor income growth. A linear factor model, including market and the innovation to the slope, explains more than half of the cross-sectional variation of average excess returns on portfolios sorted on size, book-to-market, past long term return and asset growth. The factor mimicking portfolios of the model help reconcile the joint cross section of returns on equities, equity index options, and currencies and have pricing performance comparable to several multi-factor benchmarks. My evidence suggests that the slope of the term structure is a recession state variable, and an economic source of risk premia on test assets can be attributed to time-varying investor concern over future recessions that is priced.

E0406: **Estimating cost of volatility risk in selected agricultural commodity markets**
Presenter: **Lei Yan**, University of Illinois at Urbana-Champaign, United States

The cost of volatility risk in agricultural commodity markets is investigated by examining delta-neutral straddle gains. Within a stochastic volatility model, delta-neutral straddle gains scaled by futures price are mainly determined by the price of volatility risk and its risk exposure. Using a sample of options for 2003-2016, we show that volatility risk is priced with a negative premium in the grain and livestock markets. The cost of bearing volatility risk exhibits a non-trivial term structure, with its absolute value declining sharply in maturity and approaching zero beyond three months. Regression analyses reveal that the cost of volatility risk is related to expected volatility, time to maturity, and futures trading volume, and becomes more evident on the day preceding the release of USDA reports. The results highlight the importance of volatility risk and carry important implications for option pricing and volatility forecasting in commodity markets.

EO069 Room LT-14 STATISTICAL MACHINE LEARNING METHODS AND CAUSAL INFERENCE Chair: Yingying Fan

E0359: **Using missing types to improve partial identification with application to a study of HIV prevalence in Malawi**
Presenter: **Zhichao Jiang**, Princeton University, United States

Traditional missing data analysis uses only the information of the binary missing data indicator, that is, a certain data point is either missing or not. Nevertheless, real data often provide more information than a binary missing data indicator, and they often record different types of missingness. In a motivating HIV status survey, missing data may be due to the units unwillingness to respond to the survey items or their hospitalization during the visit, and may also be due to the units temporarily absence or relocation. It is apparent that some missing types are more likely to be missing not at random, but other missing types are more likely to be missing at random. We show that making full use of the missing types results in narrower bounds of the parameters of interest. In a real life example, we demonstrate substantial improvement of more than 50% reduction in bound widths for estimating the prevalence of HIV in rural Malawi. As we illustrate using the HIV study, our strategy is also useful for conducting sensitivity analysis by gradually increasing or decreasing the set of types that are missing at random. In addition, we propose a method to construct confidence intervals for partially identified parameters with bounds expressed as the minimums and maximums of finite parameters, which is useful for not only our problem but also many other problems involving bounds.

E0608: **Optimal tradeoffs in matched designs for observational studies**
Presenter: **Samuel Pimentel**, UC Berkeley, United States
Co-authors: Rachel Kelz

An effective matched design for causal inference in observational data must achieve several goals, including balancing covariate distributions marginally, ensuring units within individual pairs have similar values on key covariates, and using a sufficiently large sample from the raw data. Yet optimizing one of these goals may force a less desirable result on another. We address such tradeoffs from a multi-objective optimization perspective by creating matched designs that are Pareto optimal with respect to two goals. We provide tools for generating representative subsets of Pareto optimal solution sets and articulate how they can be used to improve decision-making in observational study design. We illustrate the method in reanalysis of a large surgical outcomes study comparing outcomes of patients treated by US-trained surgeons and of patients treated by internationally-trained surgeons. Formulating a multi-objective version of the problem helps us evaluate the cost of balancing an important variable in terms of two other design goals, average closeness of matched pairs on a multivariate distance and size of the final matched sample.

E0673: **Combining multiple observational data sources to estimate causal effects**
Presenter: **Peng Ding**, University of California, Berkeley, United States

The era of big data has witnessed an increasing availability of multiple data sources for statistical analyses. As an important example in causal inference, we consider estimation of causal effects combining big main data with unmeasured confounders and smaller validation data with supplementary information on these confounders. Under the unconfoundedness assumption with completely observed confounders, the smaller validation data allow for constructing consistent estimators for causal effects, but the big main data can only give error-prone estimators in general. However, by leveraging the information in the big main data in a principled way, we can improve the estimation efficiencies yet preserve the consistencies of the initial estimators based solely on the validation data. The proposed framework applies to asymptotically normal estimators, including the commonly-used regression imputation, weighting, and matching estimators, and does not require a correct specification of the model relating the unmeasured confounders to the observed variables. Coupled with appropriate bootstrap procedures, our method is straightforward to implement using software routines for existing estimators.

EO220 Room LT-16 SCALABLE BAYESIAN METHODS Chair: Cheng Li

E0316: **Gaussian processes on the circle**
Presenter: **Meng Li**, Rice University, United States
Co-authors: Subhashis Ghosal

Gaussian processes indexed by the circle provide a flexible model for closed curves, an one-dimensional object in two-dimensional space. Gaussian process often has the computational hurdle in its implementation that requires repeated matrix inversion, thus may not scale well. Focusing on the squared exponential Gaussian process on the circle, we obtain the analytical eigen-decomposition of its kernel, which enables efficient posterior sampling. We further conduct an extensive study of its reproducing kernel Hilbert space. An application to boundary detection problem in image process shows that squared exponential Gaussian process on the circle guarantees the geometric restriction of the boundary, leads to nearly minimax rate estimators adaptive to the smoothness of the boundary, and is computationally efficient.

E0375: **Bayesian multi-layered Gaussian graphical models**
Presenter: **Min Jin Ha**, UT MD Anderson Cancer Center, United States

Simultaneous modeling of data arising from multiple ordered layers provides insight into the 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 develop a multi-layered Gaussian graphical model (mlGGM) to investigate conditional independence structures in probabilistic chain graphs. Our proposed model uses a Bayesian node-wise selection framework that coherently accounts for dependencies in the mlGGM. Using Bayesian variable selection strategies for each of the node-wise regressions allows for flexible modeling, sparsity and incorporation of edge-specific prior knowledge. Through simulated data generated from various scenarios, we demonstrate that our node-wise regression method outperforms other related multivariate regression-based methodologies. We apply mlGGM to identify integrative networks for key signaling pathways in kidney cancer and dynamic signaling networks using longitudinal proteomics data in breast cancer.

E0577: **Bayesian sharp minimaxity via FDR penalization**
Presenter: **Qifan Song**, Purdue University, United States

Bayesian inference is considered for high dimensional regression problems with an unknown sparse coefficient vector. In literature, various Bayesian approaches are proposed and shown to be consistent for model selection. We first study the relationship between rate minimaxity of estimation and model selection consistency, and conjecture that selection consistency estimator is supoptimal in terms of L_2 error, especially when the true coefficient is relatively dense and contains many weak signals. Inspired by the B-H FDR procedure and its minimaxity under normal means model, we propose a Bayesian modeling that corresponds to FDR and show that its corresponding posterior contraction rate is rate-minimax, and the number of false discoveries selected by posterior is bounded. More importantly, we find that under certain near orthogonal design, the posterior is asymptotically sharply minimax in terms of the multiplicative constant, and ratio of number of false discoveries over true sparsity can be arbitrarily small.

EO166 Room LT-17 MODEL UNCERTAINTY AND MODEL AVERAGE Chair: Hua Liang

E0371: **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. The exponential size of the candidate model set may also bring additional numerical error in calculating the weights. A scalable frequentist model averaging method is proposed, which is statistically and computationally efficient, to overcome this problem by using the singular value decomposition. The method enables us to find the optimal weights by considering 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., $p >> n$). Numerical studies illustrate the superiority of the proposed method in terms of both statistical efficiency and computational cost.

E0486: **Model averaging for two non-nested models**
Presenter: **Yan Gao**, Minzu University of China, China

The Mallows model averaging approach is proposed to be used for two non-nested models. It is proved that the obtained weight of true model converges to 1 with root-n rate. It develops a penalized Mallows criterion which ensures that the weight of the true model equals 1 with probability tending to 1. Simulation results indicate the consistency and also show the model averaging approach performs better than the estimation post J-test.

E0572: **Corrected Mallows model averaging approach**
Presenter: **Guohua Zou**, School of Mathematical Sciences, Capital Normal University, China

An important problem with model averaging approach is the choice of weights. The Mallows criterion for choosing weights is the first asymptotically optimal criterion, which has been used widely. We propose a corrected Mallows model averaging (MMAc) method based on small sample F distribution. MMAc exhibits the same asymptotic optimality as Mallows model averaging (MMA) in the sense of minimizing the squared errors in large sample sizes. The consistency of the MMAc based weights tending to the optimal weights minimizing MSE is also studied. We derive the convergence rate of the new empirical weights. Similar property for MMA and Jackknife model averaging (JMA) is established as well. An extensive simulation study shows that MMAc often performs better than MMA and other commonly used model averaging methods, especially for small and moderate sample size cases. The results from two real data analyses also support the proposed method.

EO222 Room LT-18 RECENT DEVELOPMENTS IN FUNCTIONAL DATA ANALYSIS Chair: Lilun Du

E0407: **Regression with dependent functional errors-in-predictors**
Presenter: **Cheng Chen**, London School of Economics, United Kingdom
Co-authors: Shaojun Guo, Xinghao Qiao

Functional regression is an important topic in functional data analysis. Traditionally, in functional regression, one often assumes that samples of the functional predictor are independent realizations of an underlying stochastic process, and are observed over a grid of points contaminated by independent and identically distributed measurement errors. However, in practice, the dynamic dependence across different curves may exist and the parametric assumption on the measurement error covariance structure could be unrealistic. We consider functional linear regression with serially dependent functional predictors, when the contamination of predictors by measurement error is "genuinely functional" with fully nonparametric covariance structure. Inspired by the fact that the autocovariance operator of the observed functional predictor automatically filters out the impact of the unobserved measurement error, we propose a novel generalized method of moments estimator of the slope parameter. The asymptotic property of the resulting estimator is established. We also demonstrate that the proposed method significantly outperforms possible competitors through intensive simulation studies. Finally, the proposed method is applied to a public financial dataset, revealing some interesting findings.

E0385: **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 country fertility and mortality data.

E0478: **Covariance and graphical modelling for high-dimensional longitudinal and functional data**
Presenter: **Cheng Qian**, London School of Economics and Political Science, United Kingdom
Co-authors: Xinghao Qiao

The problem of estimating functional covariance and graphical models from a data set consisting of multivariate sparse longitudinal data is considered. The underlying trajectories are represented through the functional principal components expansions, where the covariance matrix of principal component scores characterizes the global covariance feature and principal component functions present the functional representation of covariance relationships. Our proposed estimation procedure first implements a nonparametric method to perform functional principal components for sparse longitudinal data, and then computes functional regularized covariance or precision matrices. We derive the relevant concentration inequalities for high dimensional sparsely sampled functional data and use them to investigate the uniform consistency results for our proposed estimators. The finite sample performance of our proposed methods are illustrated through an extensive set of simulation studies and two real data examples.

EO285 Room P4302 NETWORK MODELING FOR TIME SERIES Chair: Yin Liao

E0304: **Multiple changepoint estimation in high-dimensional Gaussian graphical models**
Presenter: **Alex Gibberd**, Imperial College London, United Kingdom
Co-authors: Sandipan Roy
Many modern datasets exhibit a multivariate dependence structure that can be modelled using networks or graphs. For instance, in financial applications, one may study Markowitz minimum-variance portfolios based on sparse inverse covariance matrices. However, in reality, we expect that the underlying volatility and dependency structure of instruments may change over time, we thus require a way of tracking these dynamic network structures. We will discuss consistency properties for a regularised M-estimator which simultaneously identifies both change points and graphical dependency structure in multivariate time-series. Specifically, we will 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. Under mild conditions we present a proof of change-point consistency for this estimator. In particular, it is demonstrated that both the changepoint and graphical structure of the process can be consistently recovered, for which finite sample bounds are provided.

E0632: **A model for dynamic processes on networks**
Presenter: **Matthew Nunes**, Lancaster University, United Kingdom
Co-authors: Kathryn Leeming, Marina Knight, Guy Nason
Analysis problems are considered for time series that are observed at nodes of a potentially large network structure. Such problems commonly appear in a vast array of fields, such as environmental data or epidemiology, or measurements from computer system monitoring. The time series observed on the network might exhibit different characteristics such as nonstationary behaviour or strong correlation, and the nodal series evolve according to the inherent spatial structure. We will introduce the network autoregressive / moving average processes: a set of flexible models for network time series. For fixed networks the models are essentially equivalent to vector autoregressive moving average-type models. However, our models are especially useful when the structure of the graph, associated with the multivariate time series, changes over time. Such network topology changes are invisible to standard VARMA-like models. For integrated network time series models we introduce network differencing based on a network lifting (wavelet) transform and remark on some of its properties. We demonstrate our techniques on some real data for some example analysis talks for network time series.

E0634: **Graphical models for multivariate time series using wavelets**
Presenter: **Maria Grith**, Erasmus University Rotterdam, Netherlands
Co-authors: Matthias Eckardt
Local partial dependence and Granger causality graphs are defined for nonlinear and locally stationary multivariate time series processes using wavelet-based methods. In these graphs, nodes denote component processes, and edges describe pairwise conditional dependence between two processes, after removing the contemporaneous, lag and lead influences of the remaining variables. Local dependence is characterized by the wavelet partial coherence measures, defined in the time-frequency domain. Based on these measures, we define undirected, directed and mixed (multi)graphs, which describe specific interactions between time processes. We recover the graphs structure and the edge weights for Gaussian and non-Gaussian settings. We illustrated our methodology for simulated data and apply it to the realized volatilities of the ten largest equity indexes in the world.

EO253 Room P4704 RECENT ADVANCES FOR SEMIPARAMETRIC MODELS IN ECONOMETRICS AND STATISTICS Chair: Jingjing Wu

E0364: **Semiparametric inferences for dominance index under density ratio model**
Presenter: **Weiwei Zhuang**, University of Science and Technology of China, China
Two-sample problem is an over-investigated topic in statistics. Often, we are asked to test whether or not two populations have the same mean against one-sided or two-sided alternatives. The standard method is the t-test. However, the population means merely show the central tendency. They fail to reflect general relationship between two populations. The stochastic dominance index makes up some shortfall for the purpose of comparing two populations. We consider the problem of testing for the degree of stochastic dominance between two populations under a density ratio model (DRM). In many applications, the distributions of two populations under investigation are of similar nature. Modeling their density ratio provides an effective tool to improve the efficiency of statistical inference. We use the DRM-based empirical likelihood (EL) to estimate the stochastic dominance index and show that it is consistent and asymptotically normal. A bootstrap method is then used to construct test for the degree of dominance. Simulation study shows that this approach has superior performance than the existing method that does not activate the DRM assumption.

E0443: **Daily box office prediction model based on LSTM**
Presenter: **Yunian Ru**, Communication University of China, China
The study of movie box offices provides important support for business intelligence decision-making process, such as distribution and cinema management. The task of the daily box office prediction model is to build a dynamic prediction model to rolling forecast daily box office. It is a complex task, as the movie box office has a short life cycle, and the static data and dynamic data that affect the trend of box office are heterogeneous. LSTM recurrent neural network has the ability to memory for a long time, and it can solve the gradient vanishing and exploding problem of RNN. Modeling with LSTM can overcome the shortcoming of ARIMA, it can deal with nonlinear relations, multivariable problems. It can also overcome the shortcoming of traditional ANN models which need to specify the time dependent length. A new model of daily movie box office prediction based on LSTM is proposed. The prediction error MAPE is 30.2%. The effect of the model is better than that of the previous model. The experiment proved that the more training data collected, the better the prediction effect.

E0784: **Back to basics: Robust tests and their implications for the Prebisch singer hypothesis and economic growth**
Presenter: **Gareth Liu-Evans**, University of Liverpool, United Kingdom
Co-authors: Stephan Pfaffenzeller, Yan Zhu
The Prebisch Singer hypothesis postulates a dramatic decline in the barter terms of trade of developing countries acting as a constraint on economic development, and one prediction of this is a decline in the export earnings of developing countries. No dramatic decay in developing countries' export earnings has been observed over the most recent decades, however. The existence of a trade balance constraint effect is investigated via a recent lasso-based treatment test that can account for unknown functional form in a semiparametric model and confounding effects on the treatment variable. The unknown functional form is allowed to draw from a large number of B-spline terms, making use of the test's applicability in high-dimensional econometric settings. Moreover, a unit root-robust trend testing procedure is applied to the constituent series of an updated Grilli and Yang data set, circumventing any ambiguity over the order of integration for each series.

Authors Index