Welcome to IASC-ARS 2015

On behalf of the International organizing committee, it is our great pleasure to invite you to Singapore to attend the 9th Conference of the Asian Regional Section of the IASC (IASC-ARS 2015) to be hosted by the Department of Statistics and Applied Probability, National University of Singapore from December 17th to 19th, 2015.

The IASC-ARS conference is the main conference organized by the Asian Regional Section of the IASC every two to four years for attendees to exchange and learn about the latest information in Statistical Computing with applications. The conference topic for 2015 is "Statistical Computing: Challenges and Opportunities in Big Data Era".

IASC-ARS 2015 mission is to provide rich attendee experience by creating a forum in Big Data Era for presenting the scientific papers, exhibiting the latest studies and applications, and for networking with peers in variety areas, including but not limited to, biometrics, econometrics, data analysis, graphics, simulation, algorithms, knowledge-based systems, and Bayesian computing.

We envisage that the IASC-ARS 2015 conference will attract statistics researchers and practitioners from the academic institutes, industrial sections, and government agencies all over the world. It will provide communication channels between international experts and Asian-Pacific scientists, as well as training opportunities for graduate students and young researchers for the exchange of statistical theory, methodology, algorithm, software, and data; for the communication with statisticians and statistics practitioners; for the promotion of statistical concept to all kinds of applications.

We hope to see you in Singapore to attend the IASC-ARS 2015 conference and to experience the unique culture and beauty of Singapore. We extend our hands for a warm welcome to you all from Singapore!

Chun-houh Chen and Hock Peng Chan
Chairperson of IASC-ARS and Head of Department of Statistics & Applied Probability, National University of Singapore

Keynote Speakers

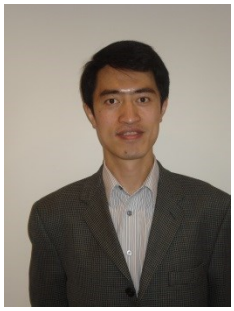
Arnaud Doucet
Department of Statistics, Oxford University, U.K.

Bio: Arnaud is Professor of Statistics at the University of Oxford. He received his Ph.D. in engineering from University of Paris XI in 1997. After completing his Ph.D., he joined Cambridge University as postdoc. He then took up permenant positions at the University of Melbourne, Cambridge University, University of British Columbia and the Institute of Statistical Mathematics. He has been professor at Oxford since 2011. His research interests include Bayesian statistics; Stochastic Simulation; Sequential Monte Carlo; Markov Chain Monte Carlo; Time Series. He is primarily interested in the development and study of novel Monte Carlo methods for inference in complex stochastic models. He is a former Canada research chair in stochastic computation at UBC and fellow of Hertford College at the University of Oxford. He has authored over 70 per-reviewed publications with over 26000 citations. He has advised over 20 doctoral students and post-docs with more than 10 holding faculty positions at various world-class universities across the world.

Yongdai Kim
Department of Statistics, Seoul National University, South Korea

Bio: Yongdai Kim is Professor of Statistics at Seoul National University. He received his Ph.D. in Statistics from Ohio State University. After completing his Ph.D., he worked at National Institues of Health as a biostatistician. In 1999, he came back to Korea and took up permanent positions are Hankuk University of Foriegn Studies and Ewha Wonmans Univesirty. IN 2004, he joined Seoul National University as an assistant professor and was promoted to full professor in 2011.
His research interests include Bayesian nonparametrics, Markov Chain Monte Carlo, Regularized methods and Variable Selection for High-dimensional Regression model and optimization algorithms.
He is currently the chair of Department of Statistics, Seoul National University. He authored over 50 peer-reviewed publications. He first proved the Bernstein-von Mises theorem for a certain semiparametric Bayesian model in 2004 and the oracle property for ultra-high dimensional models in 2008. He has supervised more than 10 doctoral students and post-docs.

Samuel Kou
Department of Statistics, Harvard University, U.S.A.



Bio: Samuel Kou is Professor of Statistics at Harvard University. He received his Ph.D. in statistics from Stanford University in 2001 under the supervision of Professor Bradley Efron. After completing his Ph.D., he joined Harvard University as an Assistant Professor. He was promoted to full professor in 2008.
His research interests include stochastic inference in single molecule biophysics, chemistry, and biology; Bayesian inference for stochastic models; nonparametric statistical methods; model selection and empirical Bayes methods; Monte Carlo methods; and economic and financial modelling.
He is the recipient of the COPSS (Committee of Presidents of Statistical Societies) Presidents' Award; a U.S. National Science Foundation CAREER Award; the Raymond J. Carroll Young Investigator Award; the Institute of Mathematical Statistics Richard Tweedie Award; and the American Statistical Association Outstanding Statistical Application Award. He is an elected Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and a Medallion Lecturer and an elected Fellow of the Institute of Mathematical Statistics.

Program

Program on 17 December 2015	
9:00-9:10	Opening Address Guest of Honour: Thorsten WOHLAND *Associate Professor, Assistant Dean (Research & Graduate Studies), Faculty of Science, National University of Singapore* Location: UTown Auditorium 2
9:10-9:15	Welcome Address Hock Peng CHAN *Professor, Head of Department of Statistics & Applied Probability, National University of Singapore* Location: UTown Auditorium 2
9:15-9:20	IASC-ARS Regional Assembly Chun-Houh CHEN *IASC-ARS Chairperson, Research Fellow and Deputy Director, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.* Location: UTown Auditorium 2
9:20–10:20	Plenary Talk 1 Chair: Patrick J.F. GROENEN *Professor of Statistics, Erasmus School of Economics* On Model Selection for Ultra-high Dimensional Models Yongdai KIM *Professor of Statistics, Seoul National University* Location: UTown Auditorium 2
10:20–10:40	Group photo and tea break
10:40–12:40	Parallel sessions 1. IS01 Smart Data in a Digital Economy (Organizer: Wolfgang Karl Härdle, Chair: Dedy Dwi Prastyo). Location: UTown Auditorium 2 2. IS09 Reliability Estimation from Degradation Data Analysis (Organizer and Chair: Bae Sukjoo) Location: LT50 3. IS12 Statistics for Environmental and Health Sciences (Organizer and Chair: Koji Kurihara) Location: LT51 4. IS27 Statistical Methods for Intractable Likelihood Functions (Organizer and Chair: Scott Sisson) Location: Seminar Room 1 5. IS16 Bayesian Analysis of High-dimensional Data (Organizer and Chair: Jaeyong Lee) Location: LT52 6. IS40 Recent Developments in Applied Econometrics and Finance (Organizer and Chair: Phuong Anh Nguyen) Location: Seminar Room 2 7. IFCS Invited session: Cluster Analysis and Multidimensional Scaling in Analysis of Marketing Data (Organizer and chair: Akinori Okada) Location: Stephen Riady Global Learning Room 8. CS01 Recent Development of Multivariate Analysis for High-dimensional Data (Organizers: Hirofumi Wakaki and Hirokazu Yanaghihara, Chair: Mariko Yamamura) Location: Seminar Room 3 9. CS02 Robust Methods in Analyzing Big Data (Organizer and Chair: Erniel B. Barrios) Location: Seminar Room 4
12:40–13:40	Lunch buffet
13:40–15:40	Parallel sessions 1. IS11 High-dimensional Classification and Regression (Organizers: Binyan Jiang and Chenlei Leng, Chair: Binyan Jiang) Location: UTown Auditorium 2 2. IS13 Cutting-edge Statistical Methods in Biomedical Sciences (Organizer and Chair: Bibhas Chakraborty) Location: LT50 3. IS39 Practical Approaches to Problems in Computational Statistics (Organizer and Chair: John Ormerod) Location: LT51 4. IS23 Systemic Risk Modeling Based on High-Frequency Data (Organizer and Chair: Sergey Ivliev) Location: LT52 5. IS34 Modeling and Analysis of Complex Biomedical Data (Organizer and Chair: Donguk Kim) Location: Stephen Riady Global Learning Room 6. IS35 Advanced Parametric and Nonparametic Statistical Approaches

	(Organizer and Chair: Yung-Seop Lee) Location: Seminar Room 1 7. IS41 Computing, Selection and Reduction in High-dimensional Statistical Methods (Organizer and Chair: Xin Zhang) Location: Seminar Room 2 8. CS03 Complex multivariate models I (Chair: Daniel Paulin) Location: Seminar Room 3 9. CS04 Models for complex biological data (Chair: Sanjay Chaudhuri) Location: Seminar Room 4
15:40–16:00	Tea break
16:00–18:00	Parallel sessions 1. IS03 Computational Econometrics and Empirical Finance (Organizer and Chair: Shih-Feng Huang) Location: UTown Auditorium 2 2. IS08 Design and Analysis for Stochastic Simulations (Organizer and Chair: Qingpei Hu) Location: LT50 3. IS17 Young Statisticians Group in IASC (YSG-IASC) Session (Organizer and Chair: Han-Ming Wu) Location: LT51 4. IS20 New Challenges in Financial Data (Organizer and Chair: Yingying Li) Location: LT52 5. IS26 Frontiers in Statistical Genomics (Organizer and Chair: Hsin-Chou Yang) Location: Stephen Riady Global Learning Room 6. IS38 Risk Analytics (Organizer and Chair: Stefan Lessmann) Location: Seminar Room 1 7. IS44 Covariance Computation (Organizer and Chair: Sanjay Chaudhuri) Location: Seminar Room 2 8. CS05 Robust Modelling and High-dimensional Data I (Chair: David Nott) Location: Seminar Room 3 9. CS06 Markov chain Monte Carlo methods (Chair: Ajay Jasra) Location: Seminar Room 4
19:00–21:00	Local tour in Singapore Night Safari (by registration)

Program on 18 December 2015	
9:00-10:00	Plenary Talk 2 Chair: Ajay JASRA *Associate Professor, Department of Statistics and Applied Probability, National University of Singapore* On a New Class of Pseudo-Marginal Algorithms Arnaud DOUCET Professor, Department of Statistics, Oxford University Location: UTown Auditorium 2
10:00-10:20	Tea break
10:20–12:20	Parallel sessions 1. IS14 Inference and Computation of Big Data in Complex Systems (Organizer and Chair: Yumou Qiu) Location: UTown Auditorium 2 2. IS24 Functional Data Analysis and Its Applications (Organizer and Chair: Ci-Ren Jiang) Location: LT50 3. IS28 Statistical Methods for Intractable Likelihood Functions (Organizer and Chair: Robert Kohn) Location: LT51 4. IS31 Circulant Orthogonal Arrays: Structure and Applications to fMRI Experiments (Organizer and Chair: Yuan-Lung Lin) Location: LT52 5. IS36 Data Representation for Analyzing Big Data (Organizer and Chair: Junji Nakano) Location: Stephen Riady Global Learning Room 6. IS37 Interpreting the Consumer: From Communication to Customization (Organizer: Tim Banks, Chair: Whye Loon Tung) Location: Seminar Room 1 7. IS45 Machine Learning and Statistical Signal Processing (Organizer: Su-Yun Huang, Chair: I-Ping Tu) Location: Seminar Room 2 8. CS07 Robust Modelling and High-dimensional Data II (Chair: Alex Beskos) Location: Seminar Room 3 9. CS08 Statistical methodology I (Chair: Fah Fatt Gan) Location: Seminar Room 4

12:20–13:30	Lunch buffet / Poster Session
13:30–15:30	Parallel sessions 1. IS04 Computation-based Statistical Process Control (Organizer and Chair: Changliang Zou) Location: UTown Auditorium 2 2. IS07 Recent Advances in Longitudinal Data Analysis (Organizer and Chair: Xingqiu Zhao) Location: LT50 3. IS10 New Developments in Financial Time Series Analysis (Organizer: Cathy W.S. Chen, Chair: Philip L.H. Yu) Location: LT51 4. IS29 Theoretical Foundation of Big Data (Organizer: Guang Cheng. Chair: Jialiang Li) Location: Stephen Riady Global Learning Room 5. ISBIS Invited Session: Innovative Statistical Methods in Business and Industry (Organizer and Chair: Yuli Hong) Location: LT52 6. CS09 Semiparametric methods (Chair: Scott Sisson) Location: Seminar Room 2 7. CS10 Complex multivariate models II (Chair: Hongmei Zhang) Location: Seminar Room 3 8. CS11 Exploratory and Graphical Methods (Chair: Alex Thiery) Location: Seminar Room 4
15:30–15:50	Tea break
15:50–17:50	Parallel sessions 1. IS05 Machine Learning in Bioinformatics and Biological Data (Organizer and Chair: Yuan-chin Ivan Chang) Location: UTown Auditorium 2 2. IS06 New Developments in Nonparametric Approaches to Analyzing High-dimensional and/or Functional Data (Organizer and Chair: Naisyin Wang) Location: LT50 3. IS15 New Trends and Approaches for High-dimensional and Complex Situations (Organizer: Yuichi Mori, Chair: Masahiro Kuroda) Location: LT51 4. IS18 High Dimension Hypothesis Testing, Change Point, Variable Selection with Categorical Variables (Organizer and Chair: Guangming Pan) Location: LT52 5. IS22 Advanced Statistical Parametric Analysis (Organizer: Jialiang Li, Chair: Binyan Jiang) Location: Seminar Room 1 6. IS30 Nature-Inspired Meta-heuristic Approaches and Their Applications in Designs of Experiments (Organizer and Chair: Frederick Kin Hing Phoa) Location: Seminar Room 2 7. IS32 Statistical Methods and Applications (Organizer and Chair: Shaoli Wang) Location: Seminar Room 3 8. IS46 Analysis and Applications of Omics Data (Organizer and Chair: Hsuan-Yu Chen) Location: Stephen Riady Global Learning Room 9. CS12 Statistical Methodology II (Chair: Dacheng Chen) Location: Seminar Room 4
19:00–22:00	Conference banquet at Chijmes Hall (by registration)

Program on 19 December 2015	
9:00–11:00	Parallel sessions 1. IS02 Theoretical and Computational Developments for Statistical Inference for Stochastic Processes (Organizer and Chair: Kengo Kamatani) Location: UTown Auditorium 2 2. IS19 Distance and Subspace Learning, Interaction Screening (Organizer and Chair: Zheng Tracy Ke) Location: LT52 3. IS21 Applications of Memetic and Metaheuristic Algorithms for Solving Biomedical Problems (Organizer and Chair: Weng Kee Wong) Location: Seminar Room 4 4. IS25 Model Specification and Selection (Organizer: I-Ping Tu, Chair: Su-Yun Huang) Location: LT50 5. IS33 Correlated Data and Financial Time Series (Organizer and Chair: Zhen Pang) Location: LT51 6. IS42 Precision Medicine: From Benchside To Bedside (Organizer and Chair: Ying Yuan) Location: Stephen Riady Global Learning Room 7. IS43 Emerging Areas in Applied Statistics (Organizer: Smarajit Bose, Chair: Chun-Houh Chen) Location: Seminar Room 1 8. CS13 Statistical modelling I (Chair: Zhou Yan) Location: Seminar Room 2
11:00-11:20	Tea break
11:20–12:20	Plenary Talk 3 Chair: Ching-Shui CHENG *Distinguished Research Fellow and Director, Academia Sinica* Big Data, Google and Disease Detection: The Statistical Story Samuel KOU *Professor, Department of Statistics, Harvard University* Location: UTown Auditorium 2
12:20–12:40	Closing Ceremony Ying CHEN *Associate Professor, Department of Statistics and Applied Probability, National University of Singapore* *Co-chair of International Organizing Committee for IASC-ARS2015* Message from the Chairperson Elect of IASC-ARS Jung Jin LEE *Professor of Statistics, Department of Statistics and Actuarial Science, Soongsil University, Korea* *Chairperson Elect of IASC-ARS* Welcome message to IASC-ARS 2017 Ciprian Doru GIURCANEANU *Senior lecturer in Statistics, Department of Statistics, University of Auckland*
12:40-14:00	Farewell lunch

Table of Contents

Welcome to IASC-ARS 2015 1
Keynote Speakers 2
Program 3
Plenary Talks 8
Invited Session 12
IS01 – IS10 13-32
IS11 – IS20 ..... 32 - 53
IS21 – IS30 ..... 54 - 72
IS31 – IS40 ..... 72 - 93
IS41 – IS46 ..... 93 -106
IFCS 107
ISBIS 109
Contributed Session 112
CS01 – CS13 ..... 113 - 140
Poster Session 141

Plenary Talks

PT01: On a New Class of Pseudo-Marginal Algorithms

Chair: Ajay JASRA
Associate Professor, Department of Statistics and Applied Probability, National University of Singapore
Venue: UTown Auditorium 2
Time: 18 Dec, 9:00-10:00

Arnaud DOUCET
Department of Statistics, Oxford University

Abstract: The use of unbiased estimators within the Metropolis—Hastings has found numerous applications in Bayesian statistics. The resulting so-called pseudo-marginal algorithm allows us to deal with intractable likelihood functions which are unbiasedly estimated using, for example, importance sampling or particle filters. However, recent theoretical results have established that the computational cost of pseudo-marginal methods is for many common applications of order T^2 for T data points at each iteration. This cost is prohibitive for large datasets. I will present new procedures which can provably significantly reduce it. On various applications, the efficiency of computations is increased by several orders of magnitude.

PT02: On Model Selection for Ultra-high Dimensional Models

Chair: Patrick J.F. GROENEN
Professor of Statistics, Erasmus School of Economics
Venue: UTown Auditorium 2
Time: 17 Dec, 9:20-10:20

Yongdai KIM
Department of Statistics, Seoul National University

Abstract: Model selection is one of the most important topics for ultra-high dimensional models where the number of covariates is much larger than the sample size. Various methods including penalized regression approached and information criteria have been proposed. In this talk, first I review what have been done so far for model selection on ultra-high dimensions, and explain what have not been done yet. Then, I will introduce recent results including fast computation and data-adaptive information criterion.

PT03: Big Data, Google and Disease Detection: The Statistical Story

Chair: Ching-Shui CHENG
Distinguished Research Fellow and Director, Academia Sinica
Venue: UTown Auditorium 2
Time: 19 Dec, 11:20-12:20

Samuel KOU
Department of Statistics, Harvard University

Abstract: Big data collected from the internet have generated significant interest in not only the academic community but also industry and government agencies. They bring great potential in tracking and predicting massive social trends or activities. We focus on tracking disease epidemics in this talk. We will discuss the applications, in particular Google Flu Trends, some of the fallacy and the statistical implications. We will propose a new model that utilizes publicly available online data to estimate disease epidemics. Our model outperforms all previous real-time tracking models for influenza epidemics at the national level of the US. We will also draw some lessons for big data applications.

Invited Session

IS01 SMART DATA IN A DIGITAL ECONOMY

Session Organizer: Wolfgang Karl Härdle, Humboldt-Universität zu Berlin, Germany
Session Chair: Dedy Dwi Prastyo, ITS Surabaya, Indonesia
Venue: UTown Auditorium 2
Time: 17 Dec, 10:40-12:40

Distillation of news flow into stock market reactions

Elisabeth Bommes
Humboldt-Universität zu Berlin, Germany

Keywords: Investor Sentiment, Attention Analysis, Sector Analysis, Volatility Simulation, Trading Volume, Returns, Bootstrap

Abstract: News carry information of market moves. The gargantuan plethora of opinions, facts and tweets on financial business offers the opportunity to test and analyze the influence of such text sources on future directions of stocks. It also creates though the necessity to distill via statistical technology the informative elements of this prodigious and indeed colossal data source. Using mixed text sources from professional platforms, blog fora and stock message boards we distill via different lexica sentiment variables. These are employed for an analysis of stock reactions: volatility, volume and returns. An increased (negative) sentiment will influence volatility as well as volume. This influence is contingent on the lexical projection and different across GICS sectors. Based on review articles on 100 S&P 500 constituents for the period of October 20, 2009 to October 13, 2014 we project into BL, MPQA, LM lexica and use the distilled sentiment variables to forecast individual stock indicators in a panel context. Exploiting different lexical projections, and using different stock reaction indicators we aim at answering the following research questions: (i) Are the lexica consistent in their analytic ability to produce stock reaction indicators, including volatility, detrended log trading volume and return? (ii) To which degree is there an asymmetric response given the sentiment scales (positive v.s. negative)? (iii) Are the news of high attention firms diffusing faster and result in more timely and efficient stock reaction? (iv) Is there a sector specific reaction from the distilled sentiment measures? We find there is significant incremental information in the distilled news flow. The three lexica though are not consistent in their analytic ability. Based on confidence bands an asymmetric, attention-specific and sector-specific response of stock reactions is diagnosed.

Digital currencies and data analytics

Ernie TEO Gin Swee
Singapore Management University, Singapore

Abstract: Since the release of Bitcoin in 2009, many forms of digital currencies have emerged in the system. They provide an interesting opportunity for data analytics as transactions on these systems are transparent and logged on the online ledger; movements of money between accounts can be identified (although these accounts are anonymous). Some digital currency systems (such as Ripple) also allow for trade between international currencies; trades and buy orders are logged in such systems. This talk will give an overview of the key digital currencies available in the market, discuss the type of data available and their potential for smart data analysis. Findings from preliminary data analysis will also be presented.

Q3, D3, LSA

Lukas Borke
Humboldt-Universität zu Berlin, Germany

Keywords: QuantNet, text mining, similarity, semantic web, document clustering, vector space model, LSA, term-term correlation, weighting scheme, visualization

Abstract: QuantNet is an integrated web-based environment consisting of different types of statistics related documents and program codes called Quantlets. The QuantMiner creates reproducibility and offers easy access by means of a powerful and specialized searching interface. This Q3-concept increases the information retrieval efficiency but there is still a need for incorporating semantic information. We employ the D3 (Data-Driven Documents) framework concentrating on semi-structured small or medium size corpora of documents. Relying on the QuantNet platform we examine 3 semantic analysis approaches: VSM (Vector Space Model), GVSM (Generalized Vector Space Model) and LSA (Latent Semantic Analysis).Amongst these three models, LSA has been successfully used for IR (Information Retrieval) purposes as a technique for capturing semantic relations between terms and inserting them into the similarity measure between two documents. Subsequently, the performance of LSA is presented applying it for the IR, document clustering and visualization tasks in the self-developed visualization framework.

Social media mining and analysis for business and consumer insights

Feida Zhu
Pinnacle Lab@Singapore Management University, Singapore

Abstract: The recent blossom of social network services has provided everyone with an unprecedented level of ease and fun of sharing information of all sorts. These public social data therefore reveal a surprisingly large amount of information about an individual which is otherwise unavailable. The business, consumer and social insights attainable from this big and dynamic social data are critically important and immensely valuable in a wide range of applications for both private and public sectors. In particular, there has been a growing interest in harnessing social media data for financial innovation. In this talk, we will explore some recent advances along this direction including personal credit scoring, risk management and customer acquisition.

IS02 THEORETICAL AND COMPUTATIONAL DEVELOPMENTS FOR STATISTICAL INFERENCE FOR STOCHASTIC PROCESSES

Session Organizer and Chair: Kengo Kamatani, Osaka University, Japan
Venue: UTown Auditorium 2
Time: 19 Dec, 9:00-11:00

Quasi likelihood analysis for ultra high frequency data

Nakahiro Yoshida
Graduate School of Mathematical Sciences, University of Tokyo
CREST JST, Japan.
E-mail: nakahiro@ms.u-tokyo.ac.jp

Keywords: quasi likelihood analysis, point process, regression, ultra high frequency data, non-ergodic statistics.

Abstract: The latest trend of financial statistics is toward analysis of ultra high frequency phenomena by modeling more precise mechanisms in a more precise time-scale. There is no Brownian motion as a driving process of the system since the central limit theorem is not effective at this level of description, differently from the standard framework. Point process modeling gives a promising approach to a description of microstructure. The quasi likelihood analysis (QLA) is a systematic analysis of the quasi likelihood random field and the associated estimators, with a large deviation method that derives precise tail probability estimates for the random field and estimators. In this talk, QLA is constructed for point processes. The point process regression model can express asynchronicity of observations and microstructure. This model can incorporate nonstationarity under finite time horizon and self-exciting/self-correcting effects of the point processes as well as exogenous effects. Non-ergodic statistics is obtained when the intensities

of the point processes become large. The point process regression model is applied to price models and limit order books. A related topic is a nonparametric method for estimating covariation between intensity processes. QLA can be developed also in long-time asysmptotics. Then establishing ergodicity of point processes becomes an issue.

Computational aspects of estimating Levy driven models

Hiroki Masuda
Kyushu University, JST CREST, Japan
E-mail: hiroki@imi.kyushu-u.ac.jp

Keywords: High-frequency sampling, Levy process, package, stochastic differential equation.

Abstract: We consider estimation problem concerning stochastic differential equations driven by a Levy process with jumps. The model is supposed to be observed at high-frequency, allowing us to incorporate a small-time approximation of the underlying likelihood. An overview of some existing theories based on the Gaussian and non-Gaussian quasi-likelihoods is presented, together with their computational aspects. Also to be demonstrated is how to implement the theory in the YUIMA package: an R framework for simulation and inference of stochastic differential equations.

Computational aspects of simulation and inference for CARMA and COGARCH models

Stefano M. Iacus
Department of Economics, Management and Quantitative Methods
University of Milan, Italy.
E-mail: stefano.iacus@unimi.it

Keywords: CARMA models, COGARCH models, inference for stochastic processes, simulation, empirical finance, yuima package

Abstract: We present a new set of tools for the R package “yuima”, available on CRAN, for the simulation and inference of Continuous Autoregressive Moving Average (CARMA) and Continuous GARCH (COGARCH) models with some applications to real data. For both CARMA (p,q) and COGARCH(p,q) models, the yuima package allows for the possibility of recovering the increments of the underlying noise via appropriate filtering. The model specification in yuima, also allows for choosing the appropriate driving Levy model for both estimation and simulation. The estimation of the parameters for the underlying Levy process makes yuima package appealing for modeling financial time series. Indeed, identifying the appropriate noise for a CARMA and COGARCH models allows to capture asymmetry and heavy tails observed in the real data. The quasi-maximum likelihood (QMLE) approach is used to estimate the parameters of the CARMA (p,q) model, while for the

COGARCH(p,q) model a mix of the generalized method of moments (GMM) and QMLE are applied. When possible, the scaling property of the Levy process is used to increase the accuracy of the estimates through aggregation of the increments.

Hybrid multi-step estimators of the volatility for stochastic regression models

Masayuki Uchida
Graduate School of Engineering Science, Osaka University, Japan
MMDS, CREST JST.
E-mail: uchida@sigmath.es.osaka-u.ac.jp

Keywords: Bayes type estimator; diffusion type processes; high frequency data; maximum likelihood type estimator

Abstract: We study the efficient estimation of the volatility parameter based on discrete observations from non-ergodic diffusion type processes on the fixed interval. Although Uchida and Yoshida (2013, SPA) proved that both the maximum likelihood (ML) and Bayes type estimators have the asymptotic mixed normality and the convergence of moments for non-ergodic diffusion type processes, from the viewpoint of numerical analysis, the optimization of the quasi likelihood function for obtaining the ML type estimators involves a serious problem, and the computation of the Bayes type estimators takes much time. Uchida and Yoshida (2012, SPA; 2014, SISP) considered adaptive ML and Bayes type estimators of both drift and volatility parameters for ergodic diffusion processes. Recently, even if an initial estimator has non-optimal rate of convergence, the multi-step estimator has the asymptotic normality of the multi-step estimator and the convergence of moments for ergodic diffusion processes, see Kamatani and Uchida (2015, SISP). It is possible to apply this method to parameter estimation for non-ergodic diffusion type processes. In this talk, we propose a hybrid multi-step estimation by using the initial Bayes type estimator with non-optimal rate of convergence. It is shown that the multi-step estimators have asymptotic mixed normality with convergence of moments. This is a joint work with Kengo Kamatani and Akihiro Nogita.

IS03 COMPUTATIONAL ECONOMETRICS AND EMPIRICAL FINANCE
Session Organizer and Chair: Shih-Feng Huang, Department of Applied Mathematics, National University of Kaohsiung, Taiwan
Venue: UTown Auditorium 2
Time: 17 Dec, 16:00-18:00

A multi-phase, flexible, and accurate lattice for pricing complex derivatives with multiple market variables

Chuan-Ju Wang
Department of Computer Science, University of Taipei, Taiwan

Abstract**:** With the rapid growth and the deregulation of financial markets, many complex derivatives have been structured to meet specific financial goals. Unfortunately, most complex derivatives have no analytical formulas for their prices, particularly when there is more than one market variable. As a result, these derivatives must be priced by numerical methods such as lattice. However, the nonlinearity error of lattices due to the nonlinearity of the derivative's value function could lead to oscillating prices. To construct an accurate, multivariate lattice, this study proposes a multiphase method that alleviates the oscillating problem by making the lattice match the "critical locations," locations where nonlinearity of the derivative's value function occurs. Moreover, our lattice has the ability to model the jumps in the market variables such as regular withdraws from an investment account, which is hard to deal with analytically. Numerical results for vulnerable options, insurance contracts guaranteed minimum withdrawal benefit (GMWB), and defaultable bonds show that our methodology can be applied to the pricing of a wide range of complex financial contracts.

The R-Squareds, active risk, and active risk volatility of hedge

Meng-Feng Yen
Department of Accountancy and Institute of Finance, National Cheng Kung University, Taiwan
E-mail: yenmf@mail.ncku.edu.tw

Keywords: Hedge Funds, R-squared, Active Risk, and Active Risk Volatility

Abstract: Prior studies, such as Titman and Tiu (2011), show that a lower multi-factor regression R-squared (i.e. higher exposure to active risk relative to systematic risk) of a hedge fund implies better future performance in the monthly return rate. Based on their results, this study aims to analyze the roles that active risk and its volatility play in analyzing the performance of low R-squared managers. Our results indicate that funds with lower active risk and lower volatility of active risk appear to outperform, on average, funds with higher active risk and higher volatility of active risk. The former appear to show superior performance, as gauged by a variety of metrics, such as the risk-adjusted return, information ratio, Sharpe ratio, and a manipulation-proof

performance measure. They also feature lower management and incentive fees than their counterparts with higher active risk and greater volatility of active risk. The rationale behind the results could be the overconfidence of hedge fund managers or the option-like characteristic of their compensation, which gives hedge fund managers an incentive to take on excess active risk.

Efficient and semi-positive definite pre-averaging realized covariance estimator

Liang-Ching Lin
National Cheng Kung University, Tainan, Taiwan

Keywords: Asynchronous trade; Dynamic filter; High dimensional high Frequency Data; Microstructure noise; Realized covariance; Semi-positive definite matrix.

Abstract: We propose an efficient and semi-positive definite (SPD) pre-averaging realized covariance estimator with the fastest convergence rate of $O_p(n^{-1/4})$. The estimator is robust to the presence of market microstructure noise and is computed with asynchronous and noisy high dimensional high frequency data. Our contributions include an innovative synchronizing technique that provides informative synchronized high frequency data without losing or distorting dependence structure, and a new correction approach that ensures semi-positive definition of realized covariance matrix without sacrificing convergence rate. Simulation study and real data analysis demonstrate superior performance compared with several alternatives.
(Joint work with Ying Chen, National University of Singapore, Singapore;
Guangming Pan, Nanyang Technological University, Singapore;
Vladimir Spokoiny, Weierstrass Institute for Applied Analysis and Stochastics, Germany.)

An EPMS price estimator for multi-asset financial derivatives

Shih-Feng Huang
Department of Applied Mathematics, National University of Kaohsiung, Taiwan
E-mail: huangsf@nuk.edu.tw

Keywords: Empirical P-martingale simulation; Esscher transform; GARCH model; Multi-asset derivatives pricing.

Abstract: This article considers the empirical P-martingale simulation (EPMS) price estimator of multi-asset financial derivatives. The corresponding change of measure process of a stochastic model is derived by the multiple dimensional Girsanov theorem or Esscher transform. For Lipschitz continuous or generic Lipschitz continuous payoff functions, the consistency of the proposed EPMS price estimator is established. In simulation study, geometric average put and maximum call options are considered under multidimensional

geometric Brownian motion or GARCH models. Numerical results indicate that the proposed price estimator is accurate and is capable of improving the efficiency of traditional Monte Carlo price estimator.

IS04 COMPUTATION-BASED STATISTICAL PROCESS CONTROL
Session Organizer and Chair: Changliang Zou, Nankai University, China
Venue: UTown Auditorium 2
Time: 18 Dec, 13:30-15:30

A state space model for multivariate Poisson count series

Nan Chen
National University of Singapore, Singapore

Abstract: This paper proposes a state space model to describe multivariate count series. The model builds on multivariate log-normal mixture of independent Poisson distribution and allows for serial dependence by considering the Poisson mean vector as a latent process driven by a nonlinear autoregressive model. In this way the model allows for a flexible cross-correlation and autocorrelation structure of the count data, and can capture the overdispersion of it as well. Monte Carlo EM algorithm together with particle filtering method provides a satisfactory estimation for the model parameters and the latent process. Based on this model, we furthermore propose a statistical process control scheme for multivariate count series. The scheme can detect general distributional changes of the process efficiently. Finally we use this model to analyze damage count series of different types collected from a power utility service area as a case study.

On-line monitoring data quality of high-dimensional data streams

Zhonghua Li
Nankai University, China

Abstract: In recent years, effective monitoring of data quality has increasingly attracted attention of researchers in the area of statistical process control (SPC). 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, we propose a conditional false discovery rate-adjusted scheme to on-line monitor the data quality of high-dimensional data streams. With thousands of input data streams, the average run length (ARL) loses its usefulness because one will likely have out-of-control (OC) signals at each time period. Hence, we first control the percentage of signals that are false alarms. Then, we compare the power of the proposed MEWMA scheme with that of two alternative methods.

Compared with two competitors, numerical results show that the proposed MEWMA scheme has higher average power.

Statistical process control for latent quality characteristics using the up-and-down test

Dongdong Xiang
East China Normal University, China

Abstract: In many applications, the quality characteristic of a product is continuous but unobservable, e.g., the critical electric voltage of electro-explosive devices. It is often of importance to monitor a process with such latent quality characteristic. The existing approaches are to specify a fixed stimulus level and test products under it to collect response outcomes (0 or 1) sequentially. Then, appropriate control charts are applied to the collected binary data sequence. However, these approaches offer limited performance. Moreover, the collected dataset under them provides little information for troubleshooting when an out-of-control signal is triggered. To overcome these limitations, this paper introduces the up-and-down test for data collection and proposes a control chart based on the test. Numerical studies show that the proposed chart is able to detect the shifts effectively, and is robust in many situations. A real example involving electro-explosive devices is given to demonstrate our proposed chart.

A distribution-free multivariate change-point model for statistical process control

Zhonghua Li
Nankai University, China

Abstract: This article develops a new distribution-free multivariate procedure 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.

IS05 MACHINE LEARNING IN BIOINFORMATICS AND BIOLOGICAL DATA
Session Organizer and Chair: Yuan-chin Ivan Chang, Academia Sinica, Taiwan
Venue: UTown Auditorium 2
Time: 18 Dec, 15:50-17:50

Genotype imputation using LD-based weighted K nearest neighbor

Chen-An Tsai
Department of Agronomy, National Taiwan University, Taiwan

Abstract: Detection of single nucleotide polymorphism (SNP) in high-throughput sequencing technologies has become efficient and robust strategies for SNP discovery and genome-wide association study. However, the conventional high-throughput genotyping techniques often produce a certain proportion of missing calls. It has been long recognized that failing to account for these missing data could dramatically reduce the power of detecting SNPs. A variety of imputation methods have been developed to impute the missing genotypes. Methods based on the K-nearest neighbors (KNN) and weighting K-nearest neighbors (wtKNN) have received some attention by considering the similarities in the haplotype structures. More recently, a number of powerful methods based on hidden Markov model (HMM) have become popular in SNPs imputation. However, these methods are time consuming or mostly suitable for small maker sets imputation and cannot exploit the structure of indirect association of tightly linked SNPs. In this study, we will propose a novel but computationally simple imputation method that is based on weighting K-nearest neighbors (wtKNN) by considering linkage disequilibrium (LD). We will demonstrate the performance of our method to impute missing SNPs using both Genotyping by sequencing (GBS) data and simulation studies. In addition, we will compare the accuracy and performance of our method with competing imputation methods.

Data-geometry based learning

Pei-Ting Chou
Department of Statistics, National Cheng-Chi University, Taiwan

Abstract: High dimensional covariate information is taken as a detailed description of any individuals involved in a machine learning and classification problem. The inter-dependence patterns among these covariate vectors may be unknown to researchers. This fact is not well recognized in classic and modern machine learning literature. In this talk, I will implement an accommodating attitude to exploit potential inter-dependence patterns embedded within the high dimensionality throughout by first computing the similarity between data nodes and then discovering pattern information in the form of Ultrametric tree geometry among almost all the covariate dimensions involved. And I will make use of these patterns to build supervised and semi-

supervised learning algorithms. My data-driven learning approaches begin with the binary-class setting, then go into the multiple-class setting. Finally, I will demonstrate the efficiencies of my learning algorithms with several datasets.

Machine learning and predictive analytics of biomedical data

James J. Chen, PhD
National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA

Abstract: Predictive analytics utilizes statistical and machine learning techniques to build models to predict outcomes and trends, and to uncover complex patterns and unexpected occurrences in data. This presentation describes the use of machine learning methods to analyze several large and complex biomedical datasets. The applications include 1) pathogen serotype/subtype identification and characterization in a Salmonella fingerprint database, 2) anomaly detection in the FDA's adverse event reporting database to identify drugs or drug classes that are associated with a particular group of adverse events, and 3) biomarker adaptive design to identify the most suitable target patient subgroup and enhance study efficiency in precision medicine. General data analytical statistical and machine learning methods involved in predictive analytics include: clustering/biclustering analysis and topic modeling for identification and characterization of subgroups, including patients, pathogens, and drug-caused adverse events, statistical tests for interaction to identify predictive biomarkers and define patient subgroups, and patient classification/prediction procedures for treatment optimization.

IS06 NEW DEVELOPMENTS IN NONPARAMETRIC APPROACHES TO ANALYZING HIGH-DIMENSIONAL AND/OR FUNCTIONAL DATA
Session Organizer and Chair: Naisyin Wang, Department of Statistics, University of Michigan, USA
Venue: LT50
Time: 18 Dec, 15:50-17:50

Detecting changes in mean functions for a functional data sequence

Yu-Ting Chen
Academia Sinica, Taiwan.
E-mail: jmchiou@stat.sinica.edu.tw, bboy0302@gmail.com

Keywords: Functional principal component analysis; Least squares; Projection; Segmentation.

Abstract: Detecting changes in the mean functions of a sequence of functional data has many applications. We propose a least squares segmentation approach to detecting multiple changes in mean functions for a functional data sequence, including the total number and the positions of the functional change-points. The least squares segmentation stage recursively detects the potential change-points for a given number, which are consistent with the correct ones if they exist. These candidates are then assured to be the genuine change-points by hypotheses testing for statistical significance. We demonstrate by simulations that the proposed approach perform reasonably well in detecting changes in the mean functions.

Extracting and integrating information from multiple longitudinal/functional data

Naisyin Wang
Department of Statistics, University of Michigan, USA
E-mail: nwangaa@umich.edu

Keywords: clustering, latent variables, model selection, spline,

Abstract: Modern medical diagnostic procedures now often involve records consisting of multiple longitudinal/functional data that reflecting certain underlying systems. We explore the use of model-selection and model-averaging approaches, with the focus of extracting different latent features in the data, to best reflect the shared and contrast information embedded in different sub-groups of subjects. Various criteria, including out-of-sample prediction, were employed to gauge the use of different types of features and the bases on which they were evaluated. Effectiveness of the new methods is demonstrated using both synthetic data and data collected through medical studies.

Bayesian nonparametric functional models for high-dimensional genomics data

Veera Baladandayuthapani,
MD Anderson, U.S.A.

Due to rapid technological advances, various types of genomic, epigenomic, transcriptomic and proteomic data with different sizes, formats, and structures have become available. These experiments typically yield data consisting of high-resolution genetic changes of hundreds/thousands of markers across the whole chromosomal map.
Modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. serial and spatial correlations). Using genome continuum models as a general principle we present a class of Bayesian methods to model these genomic profiles using functional data analysis approaches. Our methods allow for simultaneous characterization of these high-dimensional functions using non-

parametric basis functions, joint modeling of spatially correlated functional data and detection of local features in spatially heterogeneous functional data – to answer several important biological questions. We illustrate our methodology by using several real and simulated datasets and propose methods to integrate various types of genomics data as well.

The interplay between regularization and computation in high-dimensional matrix estimation

Vincent Vu,
Ohio State University, U.S.A.

The estimation of high-dimensional matrices arises naturally in multivariate problems involving the inference of pairwise relationships between many variables (or entities) based on limited samples. Examples include precision matrices, covariance matrices, Ising models, and PCA. Many estimators that have been proposed for these problems are based on penalizing the likelihood or loss, because some form of regularization is usually necessary to ensure good statistical properties. However, the computation of these estimators may not scale well with the size of the problem—typically cubic or worse time complexity. We show that in a large class of such problems, the efficient computation of these estimators depends critically on the regularization, and identify conditions where more regularization implies faster and typically parallelizable computation. Thus, regularization has both statistical and computational merits.

IS07 RECENT ADVANCES IN LONGITUDINAL DATA ANALYSIS

Session Organizer and Chair: Xingqiu Zhao, Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong, Hong Kong
Venue: LT50
Time: 18 Dec, 13:30-15:30

Robust estimation for longitudinal data with informative observation times

Xingqiu Zhao
Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong, Hong Kong

Abstract: In this paper, we focus on regression analysis of irregularly observed longitudinal data that often occur in medical follow-up studies and observational investigations. The analysis of these data involves two processes. One is the underlying recurrent event process of interest and the other is the observation process that controls observation times. Most of the existing methods, however, rely on some restrictive models or assumptions

such as the Poisson assumption. For this, we propose a class of more flexible joint models and a robust estimation approach for regression analysis of longitudinal data with related observation times. The asymptotic properties of the proposed estimators are established and a model checking procedure is also presented. The numerical studies indicate that the proposed methods work well for practical situations.

Statistics analysis of multivariate longitudinal data

Xinyuan Song
Department of Statistics, Chinese University of Hong Kong, Hong Kong

Abstract: This research considers a generalized hidden Markov model to investigate the dynamic patterns and possible heterogeneity of the associations and interrelationships among variables of interest in multivariate longitudinal data analysis. The model consists of a conditional latent variable model and a mixed hidden transition model to simultaneously address different types of dependencies within the data. The maximum likelihood procedure, coupled with the expectation-maximization algorithm and efficient sampling schemes, is developed to conduct parameter estimation. The issues of model selection and hypothesis testing are also addressed. The empirical performance of the proposed methodology is examined via simulation studies. A real data example is reported for illustration.

A Bayesian analysis of the multinomial probit model under symmetric identification

Pan Maolin
Department of Mathematics, Nanjing University, China

Abstract: Previous studies on Bayesian multinomial probit model mainly select one of the alternatives as a base category to identify the model. Some recent studies showed that the posterior predictions are sensitive to the choice of the alternative acting as the base category. We give a symmetric identification method to fully identify the model and develop a Bayesian method to analyze the model with a symmetric prior. It is shown that the predictions generated by this new method are independent of the labeling of alternatives.

Efficient estimation in semivarying coefficient models for longitudinal/clustered data

Toshio Honda
Hitotsubashi University, Japan

In semivarying coefficient models for longitudinal/clustered data, usually of primary interest is usually the parametric component which involves unknown constant coefficients. First, we study semiparametric efficiency bound for

estimation of the constant coefficients in a general setup. It can be achieved by spline regression provided that the within-cluster covariance matrices are all known, which is an unrealistic assumption. Thus, we propose an adaptive estimator of the constant coefficients when the covariance matrices are unknown and depend only on the index random variable, such as time, and when the link function is the identity function. After preliminary estimation, based on working independence and both spline and local linear regression, we estimate the covariance matrices by applying local linear regression to the resulting residuals. Then we employ the covariance matrix estimates and spline regression to obtain our final estimators of the constant coefficients. The proposed estimator achieves the semiparametric efficiency bound under normality assumption, and it has the smallest covariance matrix among a class of estimators even when normality is violated. We also present results of numerical studies. The simulation results demonstrate that our estimator is superior to the one based on working independence and so on. When applied to the CD4 count data, our method identifies an interesting structure that was not found by previous analyses.

IS08 DESIGN AND ANALYSIS FOR STOCHASTIC SIMULATIONS

Session Organizer and Chair: Qingpei Hu, Academy of Mathematics and Systems Science, Chinese Academy of Science
Venue: LT50
Time: 17 Dec, 16:00-18:00

Integrated parameter and tolerance design with computer experiments

Mei Han
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong

Abstract: Robust parameter and tolerance design are effective methods to improve process quality. Li and Wu (1999), demonstrate that the traditional two-stage approach that performs parameter design followed by tolerance design to reduce the sensitivity to variations of input characteristics is suboptimal. To mitigate the problem, they propose an integrated parameter and tolerance design (IPTD) methodology that is suitable for linear models. In this talk, a computer-aided integrated parameter and tolerance design approach for computer experiments is proposed in which the means and tolerances of input characteristics are simultaneously optimized to minimize the total cost. A Gaussian process metamodel is used and a multiobjective optimization approach is proposed to find robust optimal solutions.

Pairwise meta-modeling of multivariate output computer models using nonseparable covariance function

Li Yongxiang
Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong

Abstract: The Gaussian process (GP) model is a popular method for emulating deterministic computer simulation models. Its natural extension to computer models with multivariate outputs employs a multivariate Gaussian process (MGP) framework. Nevertheless, with significant increase in the number of design points and the number of model parameters, building a MGP model is a very challenging task. Under a general MGP model framework with nonseparable covariance functions, we propose an efficient meta-modeling approach featuring a pairwise model building scheme. The proposed method has excellent scalability even for a large number of output levels. Some properties of the proposed method have been investigated and its performance has been demonstrated through several numerical examples.

Modelling regression quantile process using monotone B-splines

Nan Chen
Department of Industrial & Systems Engineering, National University of Singapore, Singapore

Abstract: Quantile regression as an alternative to conditional mean regression (i.e., least square regression) is widely used in many areas. It can be used to study the covariate effects on the entire response distribution by fitting quantile regression models at multiple different quantiles or even fitting the entire regression quantile process. However, estimating the regression quantile process is inherently difficult because the induced conditional quantile function needs to be monotone at all covariate values. In this paper, we proposed a regression quantile process estimation method based on monotone B-splines. The proposed method can easily ensure the validity of the regression quantile process, and offers a concise framework for variable selection and adaptive complexity control. We thoroughly investigated the properties of the proposed procedure, both theoretically and numerically. We also use a case study on wind power generation to demonstrate its use and effectiveness of the proposed method.

A stochastic expectation-maximization algorithm for the analysis of system lifetime data with known signature

Tony Ng
Southern Methodist University, USA

Abstract: Statistical estimation of the model parameters of component lifetime distribution based on system lifetime data with known system structure is

discussed here. We propose the use of stochastic expectation-maximization (SEM) algorithm for obtaining the maximum likelihood estimates of model parameters based on complete and censored system lifetimes. Different ways of implementing the SEM algorithm are also studied. We have shown that the proposed methods are feasible and are easy to implement for various families of component lifetime distributions. The proposed methodologies are then illustrated with two popular lifetime models – the Weibull and Birnbaum-Saunders distributions. Monte Carlo simulation is then used to compare the performance of the proposed methods with the corresponding estimation by direct maximization.

IS09 RELIABILITY ESTIMATION FROM DEGRADATION DATA ANALYSIS

Session Organizer and Chair: Bae Sukjoo, Hanyang University, South Korea
Venue: LT50
Time: 17 Dec, 10:40-12:40

Accelerated degradation test planning with product initial performance incorporated

Qingpei Hu
Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China

Abstract: In accelerated degradation test planning products are usually assigned to different stress levels randomly when the sample proportion of each level is determined, i.e. the unit to unit difference is not taken into account. However, under some circumstances, the difference from unit to unit or from batch to batch cannot be neglected. Samples with different initial performance behave differently under the same stress level during the testing, samples with the same initial performance will perform differently in different stress levels. For each sample, there is a best stress level for it to work. In this work, the initial performance information is incorporated to plan the degradation test. The Mixed-Gaussian distribution is used to describe the initial performance information and the normal random effect model is explored to illustrate our approach, and comparison is also conducted with traditional approach.

Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data

Yili Hong
Department of Statistics, Virginia Tech, Blacksburg, VA, 24061, USA

Abstract: Degradation data provide a useful resource for obtaining reliability information for some highly reliable products and systems. In addition to

product/system degradation measurements, it is common nowadays to dynamically record product/system usage as well as other life-affecting environmental variables such as load, amount of use, temperature, and humidity. We refer to these variables as dynamic covariate information. In this paper, we introduce a class of models for analyzing degradation data with dynamic covariate information. We use a general path model with individual random effects to describe degradation paths and a vector time series model to describe the covariate process. Shape restricted splines are used to estimate the effects of dynamic covariates on the degradation process. The unknown parameters in the degradation data model and the covariate process model are estimated by using maximum likelihood. We also describe algorithms for computing an estimate of the lifetime distribution induced by the proposed degradation path model. The proposed methods are illustrated with an application for predicting the life of an organic coating in a complicated dynamic environment (i.e., changing UV spectrum and intensity, temperature, and humidity). This paper has supplementary material online.

Reliability evaluation techniques in the fuel cell technology

Suk Joo Bae
Department on Industrial Engineering, Hanyang University, Korea

Abstract: Fuel cells (FCs) have received much attention as remarkable alternatives to current battery technologies for portable electronic devices and electro-mobiles. The state-of-art FCs have had a difficulty in commercialization in terms of reliability and cost. To improve reliability of FCs, I will present degradation models for FCs under various environment conditions and reliability prediction methods through accelerated degradation testing in this seminar. Reliability evaluation techniques are mainly based on statistical degradation modeling of the FCs. New reliability issues for carbon nanotubes, as a prominent material for fuel cells, are also discussed.

IS10 NEW DEVELOPMENTS IN FINANCIAL TIME SERIES ANALYSIS
Session Organizer: Cathy W.S. Chen, Feng Chia University, Taiwan
Session Chair: Philip L.H. Yu, the University of Hong Kong, Hong Kong
Venue: LT51
Time: 18 Dec, 13:30-15:30

Data-driven particle filters for particle Markov chain Monte Carlo

Catherine Forbes
Monash University, Australia

Abstract: This paper proposes new automated proposal distributions for sequential Monte Carlo (SMC) algorithms, including particle filtering and related sequential importance sampling methods. The weights for these

proposal distributions are easily established, as is the unbiasedness property of the resultant likelihood estimators, so that the methods may be used within a particle Markov chain Monte Carlo (PMCMC) inferential setting. Simulation exercises, based on a range of important financial models, are used to demonstrate the linkage between the signal-to-noise ratio of the system and the performance of the new particle filters, in comparison with existing filters. In particular, we demonstrate that one of our proposed filters performs well in a high signal-to-noise ratio setting, that is, when the observation is informative in identifying the location of the unobserved state. A second filter, deliberately designed to draw proposals that are informed by both the current observation and past states, is shown to work well across a range of signal-to-noise ratios and to be much more robust than the auxiliary particle filter, which is often used as the default choice. We then extend the study to explore the performance of the PMCMC algorithm using the new filters to estimate the likelihood function, once again in comparison with existing alternatives. The comparison is based on the optimal computing time required to estimate the posterior distribution of the parameter of interest.

Statistical estimation for optimal dividend barrier with insurance portfolio

Hiroshi Shiraishi
Department of Mathematics, Keio University, Yokohama, Kanagawa, Japan
E-mail: shiraishi@math.keio.ac.jp

Keywords: Ruin theory; Dividend; Compound Poisson process; Statistical Estimation

Abstract: We consider a problem where an insurance portfolio is used to provide dividend income for the insurance company's shareholders. This is an important problem in application of risk theory. Whenever the surplus attains the level barrier, the premium income is paid to the shareholders as dividends until the next claim occurs. In this presentation, we consider the classical compound Poisson model as the aggregate claims process, under which the dividends are paid to the shareholders according to a barrier strategy. Optimal dividend barrier is defined as the level of the barrier that maximizes the expectation of the discounted dividends until ruin, which was initially proposed by De Finetti (1957) in the discrete time model and thereafter discussed by Buhlmann (1970) in the classical risk model with the foundation laid. Although this problem was actively studied around 2000, it is less discussed with the estimation of the unknown barrier. In this presentation, we propose the estimation problem of the optimal dividend barrier, which is critical in application. We establish an estimator of the expected discounted dividends and the estimated optimal dividend barrier as its maximizer, which are shown to be consistent. Numerical simulation experiments demonstrate that the proposed estimators work well with reasonable size of samples.

Autoregressive conditional negative Binomial model applied to over-dispersed time series of counts

Cathy W. S. Chen
Feng Chia University, Taiwan
E-mail: chenws@mail.fcu.edu.tw

Abstract: In recent years, there has been growing interest in studying integer-valued time series. To accommodate time-varying over-dispersion, we propose a new model for time series of count: the autoregressive conditional negative binomial (ACNB) model. The location and scale parameters of the negative binomial distribution are flexible in the ACNB set-up where we allow time-varying conditional mean and conditional variance to handle dynamic over-dispersion. We adopt Bayesian methods with a Markov chain Monte Carlo sampling scheme to estimate model parameters. We utilize deviance information criterion for model comparison. We conduct simulations to investigate the estimation performance of the proposed negative binomial model. To demonstrate the proposed approach in modeling time-varying over-dispersion, we consider New South Wales (NSW) crime data sets. We also fit the autoregressive conditional Poisson model to these two datasets. Our results demonstrate that the proposed negative binomial model is preferable to the Poisson model.

IS11 HIGH-DIMENSIONAL CLASSIFICATION AND REGRESSION

Session Organizer: Binyan Jiang, Hong Kong Polytechnic University, Hong Kong, and Chenlei Leng, University of Warwick, UK
Session Chair: Binyan Jiang, Hong Kong Polytechnic University
Venue: UTown Auditorium 2
Time: 17 Dec, 13:40-15:40

Multiclass sparse discriminant analysis

Qing Mai
Department of Statistics, Florida State University, USA

Abstract: In recent years many sparse linear discriminant analysis methods have been proposed for high-dimensional classification and variable selection. However, most of these proposals focus on binary classification and they are not directly applicable to multiclass classification problems. There are two sparse discriminant analysis methods that can handle multiclass classification problems, but their theoretical justifications remain unknown. In this talk, we propose a new multiclass sparse discriminant analysis method that estimates all discriminant directions simultaneously. We show that when applied to the binary case our proposal yields a classification direction that is equivalent to those by two successful binary sparse LDA methods in the literature. An efficient algorithm is developed for computing our method with high-dimensional data. Variable selection consistency and rates of convergence

are established under the ultrahigh dimensionality setting. We further demonstrate the superior performance of our proposal over the existing methods on simulated and real data.

Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening

Rui Pan
School of Statistics and Mathematic, Central University of Finance and Economics, China

Keywords: Multi-class linear discriminant analysis; Pairwise sure independence screening; Sure independence screening; Strong screening consistency.

Abstract: This paper is concerned with the problem of feature screening for multi-class linear discriminant analysis under ultrahigh dimensional setting. We allow the number of classes to be relatively large. As a result, the total number of relevant features is larger than usual. This makes the related classification problem much more challenging than the conventional one, where the number of classes is small (very often two). To solve the problem, we propose a novel pairwise sure independence screening method for linear discriminant analysis with an ultrahigh dimensional predictor. The proposed procedure is directly applicable to the situation with many classes. We further prove that the proposed method is screening consistent. Simulation studies are conducted to assess the finite sample performance of the new procedure. We also demonstrate the proposed methodology via an empirical analysis of a real life example on handwritten Chinese character recognition.

On dimensionality effects in linear discriminant analysis for large dimensional data

Cheng Wang
Department of Mathematics, Shanghai Jiao Tong University, China

Keywords: Large dimensional data; linear discriminant analysis; Random matrix theory; asymptotic distribution.

Abstract: We study the asymptotic results of linear discriminant analysis (LDA) in large dimensional data where the observation dimension p, is of the same order of magnitude as the sample size n. Roughly, we know when $\frac{\mathrm{p}}{\mathrm{n}} \to 0$, LDA is an "good" classifier which means the empirical misclassification error tends to the theoretical one and if $\frac{\mathrm{p}}{\mathrm{n}} \to \mathrm{y} \in (0,1)$, we should pay some price for estimating the means and covariance matrix. The explicit theoretical results about dimensionality effects in LDA will be derived in this work. Specially, we get the asymptotic distribution of the misclassification error using recent results in random matrix theory. Based on these results, a scale

adjusted classifier will be suggested to the classical LDA to handle data with un-equal sample sizes. Finally, simulations will be conducted to support these results.

Sufficient forecasting using factor models

Lingzhou Xue
Department of Statistics, Pennsylvania State University, USA

Abstract: We consider forecasting a single time series when there is a large number of predictors and a possible nonlinear effect. The dimensionality was first reduced via a high-dimensional factor model implemented by the principal component analysis. Using the extracted factors, we develop a link-free forecasting method, called the sufficient forecasting, which provides several sufficient predictive indices, inferred from high-dimensional predictors, to deliver additional predictive power. Our method is also applicable to cross-sectional sufficient regression using extracted factors. The connection between the sufficient forecasting and the deep learning architecture is explicitly stated. The sufficient forecasting correctly estimates projection indices of the underlying factors even in the presence of a nonparametric forecasting function. The proposed method extends the sufficient dimension reduction to high-dimensional regimes by condensing the cross-sectional information through factor models. We derive asymptotic properties for the estimate of the central subspace spanned by these projection directions as well as the estimates of the sufficient predictive indices. We also show that the natural method of running multiple regression of target on estimated factors yields a linear estimate that actually falls into this central subspace. Our method and theory allow the number of predictors to be larger than the number of observations. We finally demonstrate that the sufficient forecasting improves upon the linear forecasting in both simulation studies and an empirical study of forecasting macroeconomic variables.

IS12 STATISTICS FOR ENVIRONMENTAL AND HEALTH SCIENCES
Session Organizer and Chair: Koji Kurihara, Okayama University, Japan
Venue: LT51
Time: 17 Dec, 10:40-12:40

Spatial hierarchical modeling of thermoluminescent dosimetry measurements in tiles and bricks related to the Hiroshima atomic bomb

Harry M. Cullings
Departments of Statistics, Radiation Effects Research Foundation, Hiroshima and Nagasaki, Japan
Email: hcull@rerf.or.jp

Keywords: spatial hierarchical model, geographically weighted regression

Abstract: Tiles and bricks from buildings in Hiroshima and Nagasaki were measured by various investigators starting in the 1960s, to estimate the radiation they received from the atomic bombs, using thermoluminescent dosimetry. The measurements were used to validate dosimetry systems that calculate doses received by survivors of the bombs, based on location and shielding. Recently, some investigators have claimed that spatial patterns exist in the measurements, different from the circular symmetry of the calculated doses, which they attribute to residual radioactivity that persisted in the environment after the bombings. This work analyzed 117 measurements taken at 26 sites in Hiroshima. First, geographically weighted regression (GWR) was used to relate measured values (M) to those calculated (C) by the dosimetry system, indicating a large deficit (M<C) in one small area near the bomb hypocenter and a smaller, more widely distributed excess in other areas. Then a spatial hierarchical model (SHM) was constructed to allow more detailed and specific modeling of known quantities and relationships, particularly to distinguish uncertainty that is multiplicative (terms proportional to measured dose) vs. additive (terms related to accumulated dose from background). Spatial patterns in the results of the SHM will be contrasted to those of the GWR and the raw data and possible causes will be discussed.

Detection of cluster for radiation monitoring data based on scan statistic

Fumio Ishioka
The Graduate School of Environmental and Life Science, Okayama University, Japan
E-mail: shioka@okayama-u.ac.jp

Keywords: spatial scan statistic; echelon analysis; radiation dose rate

Abstract: The Tokyo Electric Power Fukushima I Nuclear Power Plant accident released large amounts of radioactive materials to the environment. Measurement of radiation dose rate by monitoring post is carrying out by related ministries and agencies, local governments, nuclear operator and related companies in real time. This study was performed to find out the time and the location for the high contaminant clusters among the "difficult-to-return zone" in Fukushima prefecture from these monitoring results. We apply the spatial scan statistic, which performs a scan procedure for spatial data and obtains a likelihood ratio test statistic, to detect some clusters consisting of high rate radiation dose. Here, it is very important to select a scanning method properly. Some scanning approaches are proposed so far, but most of them are limited in the shape of a detected cluster, or need an unrealistic computational time if the data size is too large. To solve these problems, we have proposed to use an echelon analysis as the scanning method. The echelons enable detecting the clusters consisting of the various shapes which have high-likelihood, because the areas are scanned based on the inherent hierarchical structure of data. We apply the echelons to the monitoring post data and compare it to other scanning methods. In addition, we verify a trend of the movements of radioactive materials released in the environment.

A novel approach for comparing spatial data obtained by different measurement systems

Makiko Oda
National Defense Medical College, Japan
E-mail: oda@ndmc.ac.jp

Keywords: spatial data, hierarchical structure, habitat suitability, patch.
Abstract: A patch is defined as a spatially homogeneous area where at least one variable has similar attributes either of category or quantitative value (Fortin et al., 2005). Patches can be identified using various variables such as tree or animal abundance and percentage coverage of trees. Oda et al. (2012) suggested the technique for identification of patches in a forest using Echelon analysis. Laser and sonar are interferometric measuring systems used in river geomorphologic surveying. Accuracy and capabilities of these techniques differ which makes them suitable for different uses. We used both methods to collect river topography and further identified juvenile salmon patches based on the suitability index by habitat modeling. Two separate models (with and without detail laser data) were produced. We then compared suitable patches formed by habitat model by statistical patch analysis. Significance of accurate components used for modeling was clearly seen in patch configuration.

Pareto-type distributions on the cylinder

Kunio Shimizu,
The Institute of Statistical Mathematics, Japan
E-mails: k-shmz@ism.ac.jp

Keywords: Directional statistics; Gamma mixture; Generalized Pareto distribution; Regression.

Abstract: A point on a cylinder may be viewed as the intersection of a line and a circle, and a random vector on the cylinder in directional statistics consists of a linear variable and an angular one. Examples of cylindrical observations include wind speed and wind direction, concentration of pollutants and wind direction in environmental science, and distance and direction in the areas of animal movement and earthquake. For linear data, Pareto distributions are widely used in hydrology and climatology for expressing positive and long-tailed distributions. In this paper, we propose a Pareto-type distribution on the cylinder using a gamma mixture to study further the Johnson and Wehrly model. The marginal distributions are a wrapped Cauchy for the circular variable and an unfamiliar distribution for the linear variable. The conditional distribution of the linear variable given the circular variable is a generalized Pareto distribution, whose conditional mean provides the regression function of the circular variable. An alternative cylindrical model whose conditional distribution of the linear variable given the circular variable is the same generalized Pareto distribution is also introduced.

IS13 CUTTING-EDGE STATISTICAL METHODS IN BIOMEDICAL SCIENCES

Session Organizer and Chair: Bibhas Chakraborty, National University of Singapore, Singapore
Venue: LT50
Time: 17 Dec, 13:40-15:40

Design of sequential multiple assignment randomized trial (SMART) with ordinal outcome

Palash Ghosh
Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, Singapore

Abstract: Sequential multiple assignment randomized trials (SMART) are used to develop optimal treatment strategies for patients based on their medical histories in different branches of medical science like behavioral science and oncology, where a sequence of treatments are given to the patients. In the existing literature, SMART study had been conducted assuming final outcome as a continuous variable. However, the final outcome could be ordinal also, for example, the final outcome in a SMART as toxicity level (mild, moderate, severe) is an ordinal variable. In this work, we design SMART for ordinal outcome. More specifically, we compare two regimes (sequence of drugs) that start with different initial drugs. The required sample size formula has been derived. We discuss the issues related to response rates corresponding to two regimes. A simulation study shows estimated power corresponding to the derived sample size formulae. Key words: Ordinal outcome, SMART design, sample size, response-rate.

A powerful approach to estimation of intervention effects on fold-increase endpoints using paired interval-censored data

Xu Ying, PhD;
Assistant Professor, Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, National University of Singapore, Singapore

Abstract: Many epidemiological investigations involve laboratory assay data that are interval censored. An example is the hemagglutination inhibition (HI) assay, which measures influenza antibody concentration. Given a pair of such measurements from the same subject at two time points, a binary 'fold-increase' endpoint can be defined according to the ratio of these two measurements, as it often is in vaccine clinical trials. The intervention effect can be assessed by comparing the binary endpoint between groups of subjects given different vaccines or placebos. Conventional approaches disregard the interval-censoring nature in the measurement and analyze them as if they were continuous, which inevitably lead to reduced statistical efficiency and power loss. This talk will introduce a parametric approach to the analysis of such data in the estimation of intervention effect. Simulation study

demonstrates that this alternative approach is shown to be more statistically powerful than conventional approach. Data from two influenza trials will be used to illustrate.

An empirical Bayes approach to integrate multiple GWAS with gene expressions from multiple tissues

Jin Liu, PhD
Assistant Professor, Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, National University of Singapore, Singapore

Abstract: To date, a large number of genome-wide association studies (GWAS) have been conducted. With advancement of array techniques, there are a large number of genomic data available from multiple sources: e.g., gene expression data from multiple tissues. The unbiased tissue studies can reveal new dimensions of biological effects [Dermitzakis, 2012]. Thus, it becomes essential to integrate gene expression from multiple tissues with GWAS that can increase the statistical power of the analysis of a single GWAS. We propose to use empirical-Bayes-based approach to model the status of each gene (null or non-null) enhanced by gene expression from tissues. We develop an expectation-maximization (EM) algorithm to optimize the corresponding complete log-likelihood function. These methods can jointly analyze two or more GWAS at the same time to test for the ``pleiotropic'' effects. We can also evaluate the significance of integrating a tissue. To integrate multiple tissues, we propose a three-stage strategy using penalized linear discriminant analysis (LDA) to transform expressions from multiple tissues to the new predictor with much lower dimension. Meanwhile, we estimate the corresponding local false discovery rate (FDR) and formulate the hypothesis testing for "pleiotropy" and identification of associated tissues. Simulation studies are used to evaluate finite sample performance. We make comparison under different level of ``pleiotropy'' using generative model. Rheumatoid arthritis and type-1 diabetes from the Wellcome Trust Case Control Consortium (WTCCC) together with gene expression from multiple tissues are analyzed using the proposed approach.

Bayesian evidence synthesis for public health modelling

Alex Cook
National University of Singapore, Singapore

Abstract: Modelling is increasingly moving out of mathematics and computer science departments and into the realms of public health research and practice. For it to inform policy in an evidence-based fashion, modelling needs to be rigorously data-driven, and frequently involves pooling data from multiple disparate sources. One mechanism to do that is to use Bayesian techniques to synthesise evidence from multiple sources. In this talk I will illustrate Bayesian evidence synthesis via case studies in public health.

IS14 INFERENCE AND COMPUTATION OF BIG DATA IN COMPLEX SYSTEMS
Session Organizer and Chair: Yumou Qiu, University of Nebraska Lincoln, USA
Venue: UTown Auditorium 2
Time: 18 Dec, 10:20-12:20

Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating

Hui Huang,
Peking University, China

Keywords: Air Quality; Meteorological Condition; Observational Study; Quasi-Experiment

Abstract: By learning the PM2.5 readings and meteorological records in 2010-2015, the severity of PM pollution in Beijing is quantified with a set of statistical measures. As PM2.5 concentration is highly influenced by meteorological conditions, we propose a statistical approach to adjust PM2.5 concentration with respect to meteorological conditions, which can be used to monitor PM2.5 pollution in a location. The adjusted monthly averages and percentiles are employed to test if the PM2.5 levels in Beijing have been lowered since the China's State Council set up a pollution reduction target. The results of the testing reveal significant increases, rather than decreases, in the PM2.5 concentrations in the years 2013 and 2014 as compared to those in year 2012, respectively. We conduct analyses on two quasi-experiments: the Asia-Pacific Economic Cooperation (APEC) meeting in November 2014 and the annual winter heating, to gain insight on the impacts of emissions on PM2.5. The analyses lead to a conclusion that a fundamental shift from the mainly coal-based energy consumption to much greener alternatives in Beijing and the surrounding North China Plain is the key to solving the PM2.5 problem in Beijing.

A scalable integrative model for heterogeneous genomic data types under multiple conditions

Yingying Wei
Department of Statistics, The Chinese University of Hong Kong.
E-mail: ywei@sta.cuhk.edu.hk

Keywords: Genomics; Hierarchical model; EM algorithm; Scalable.

Abstract: A key problem in biology is how the same copy of a genome within a person can give rise to hundreds of cell types. Plentiful convincing evidence indicates multiple elements, such as transcription factor binding, histone modification, and DNA methylation, all contribute to the regulation of gene expression levels in different cell types. Therefore, it is crucial to understand

how these heterogeneous regulatory elements collaborate together, how the cooperation at a given genomic region changes across diverse cell lines, as well as how such dynamic cooperation patterns across cell lines vary along the whole genome. Here, we propose a scalable hierarchical probabilistic generative model to cluster genomic regions according to the dynamic changes of their open chromatin and DNA methylation status across cell types. The model will overcome the exponential growth of parameter space as the number of cell types integrated increases. The fitted results of the model will provide a genome-wide region-specific, cell-line-specific open chromatin and DNA methylation landscape map.

Bootstrap tests on high dimensional covariance matrices with applications to understanding gene clustering

Wen Zhou
Department of Statistics, Colorado State University, U.S.A.
riczw@stat.colostate.edu

Abstract: Recent advancements in genomic study and clinical research have drew growing attention to understanding how relationships among genes, such as dependencies or co-regulations, vary between different biological states. Complex and unknown dependency among genes, along with the large number of measurements, imposes methodological challenge in studying genes relationships between different states. Starting from an interrelated problem, we propose a bootstrap procedure for testing the equality of two unspecified covariance matrices in high dimensions, which turns out to be an important tool in understanding the change of gene relationships between states. The two-sample bootstrap test takes maximum advantage of the dependence structures given the data, and gives rise to powerful tests with desirable size in finite samples. The theoretical and numerical studies show that the bootstrap test is powerful against sparse alternatives and more importantly, it is robust against highly correlated and nonparametric sampling distributions. Encouraged by the wide applicability of the proposed bootstrap test, we design a gene clustering algorithm to understand gene clustering structures. We apply the bootstrap test and gene clustering algorithm to the analysis of a human asthma dataset, for which some interesting biological implications are discussed.
(Joint work with Jinyuan Chang and Wen-xin Zhou.)

Big data analysis in National College Entrance Examinations of China

Zili Zhang
Southwest University, Chongqing, China.

Abstract: NCEE (Gaokao in Chinese) – National College Entrance Examination in China – is a national wide examination sitten by millions of students each year. One of the core steps in Gaokao is the marking of

examination papers. In the network-based marking process, a large amount of data was collected.
In this talk, I'll discuss how to analyze the data in real time to surveil the marking process based on more than 10 year practice. I'll also discuss how to predict and measure students' abilities through data analysis. Finally, some issues and challenges will be explored.

IS15 NEW TRENDS AND APPROACHES FOR HIGH-DIMENSIONAL AND COMPLEX SITUATIONS
Session Organizer: Yuichi Mori, Okayama University of Science, Japan
Session Chair: Masahiro Kuroda, Okayama University of Science, Japan
Venue: LT51
Time: 18 Dec, 15:50-17:50

Information theoretic criteria for least-squares trees

Ciprian Doru Giurcaneanu
Department of Statistics, University of Auckland, New Zealand
E-mail: c.giurcaneanu@auckland.ac.nz

Keywords: Phylogenetic Trees; Stochastic Complexity; Minimum Description Length Principle.

Abstract: Identifying the correct evolutionary tree is an essential and difficult biological problem. It is important to neither over-resolve nor falsely resolve its structure; a problem well-suited to information criteria. Stochastic Complexity (SC) was introduced in [Rissanen(1978)] and since then various forms of it have been derived (see [Rissanen(2012)] for the newest developments of this topic). According to the MDL principle, SC is defined in the context of transmitting the existing data to a hypothesized decoder. The "encoding" is performed by using mathematical models that belong to a pre-defined class, and the model which leads to the shortest code length is deemed to be the most suitable for describing the data [Grünwald(2007)]. In this work, we consider SC for assessing phylogenetic trees. To this end, we use SC to encode the parameters and the model (tree) structure. We perform a theoretical comparison of SC with the well-known Bayesian Information Criterion (BIC) and investigate their behavior when the size of the tree$\to\infty$and as error $\to$ 0. Experiments are conducted with real-world and simulated data in which we compare SC with various forms of BIC, AIC (Akaike Information Criterion) and KIC (Kullback Information Criterion).

Penalized estimation of high-dimensional covariance matrix via matrix-logarithm transformation

Philip L.H. Yu
Department of Statistics & Actuarial Science, the University of Hong Kong, Hong Kong
E-mail: plhyu@hku.hk

Keywords: Covariance matrix estimation; Matrix-logarithm transformation; Penalization.

Abstract: It is well known that in estimating the covariance matrix Σ from a p-dimensional data set, the sample covariance matrix S is not a good estimator of Σ when the dimension p becomes large. Typically, the eigenvalues of S will be distorted. Many methods have been proposed to tackle this problem, for instance regularizing Σ by thresholding or banding. In this paper, we first approximate the data likelihood using the matrix-logarithm transformation of Σ (denoted by A) and then estimate Σ by imposing the penalty $||\mathrm{A} - \mathrm{mI}||_F^2$ to the matrix-logarithm transformed likelihood function so that the estimated eigenvalues of A are shrunk toward its mean m. Our proposed method guarantees that the estimate of Σ is always non-negative definite and the estimation is computationally efficient. The simulation study and two real data examples on portfolio optimization and classification of genomic data show the proposed method outperforms some of the existing methods. This is a joint work with Anita Wang.

Exploratory symbolic data analysis with dimension reduction methods

Han-Ming Wu
Department of Mathematics, Tamkang University, Taiwan
E-mail: hmwu@mail.tku.edu.tw

Keywords: Big data; Data visualization; exploratory data analysis; Symbolic data analysis; Principal components analysis; sliced inverse regression.

Abstract: Exploratory data analysis (EDA) serves as a preliminary yet essential tool for summarizing the main characteristics of a data set before appropriate statistical modeling can be applied. Quite often, EDA employs the traditional graphical techniques such as the boxplot, histogram and scatterplot and are equipped with various dimension reduction methods and computer-aided interactive functionalities. EDA has been used to explore different data types. Examples were the cases of the survival data, the time series data, the functional data and the longitudinal data. Conventionally, these data set were tabulated by a table with p columns corresponding to p variables. Each subject is measured by a single numerical value for each variable. Nowadays the collected data keeps getting much bigger and more complex. The description of data was no longer stored by a form of a single value but the intervals, histograms and/or distributions. These are examples of the so-called

symbolic data. This study intends to develop EDA with more visual methods for symbolic data. Two dimension reduction methods, the principal component analysis (PCA) and the sliced inverse regression (SIR), are also extended and used to reveal the insight structure of symbolic objects embedded in the high-dimensional space. On the other hand, the statisticians are facing the challenges of analyzing the big data that are gathered rapidly from diverse resources with complex types. SDA supplies various data descriptions and has great capacity for big data. As a consequence, exploratory symbolic data analysis (ESDA) as a tool that supports the efficient, effective and practical exploration of symbolic data sets is needed.

Acceleration of the alternating least squares algorithm for nonlinear multivariate analyses

Yuichi Mori
Okayama University of Science, Okayama, Japan
E-mail: mori@soci.ous.ac.jp,

Keywords: vector ε algorithm; nonlinear PCA; nonlinear FA; mixed measurement level data; simulation study.

Abstract: The alternating least squares (ALS) algorithm is often utilized in nonlinear multivariate analyses for mixed measurement level data, where optimal transformation and low-rank matrix approximation are alternated until convergence. Since the ALS algorithm requires many iterations and much computation time for convergence due to its linear convergence, we have proposed an acceleration algorithm to speed up ALS computation using the vector “(vε-ALS) algorithm in nonlinear principal component analysis (NL-PCA) and nonlinear factor analysis (NL-FA). We apply the proposed procedures to more complex situations such as variable selection problem and large datasets with a variety of measurement levels, and evaluate how much the algorithm improves computational efficiency in such cases. We conduct some experiments in which NL-PCA and NL-FA using the vε-ALS algorithm are applied to a couple of real datasets and several artificial datasets which have large numbers of variables with a variety of mixing rates of numerical and categorical variables. Those experiments indicate that the performance of vε-ALS algorithm in both NL-PCA and NL-FA is improved about 3 times of ordinary ALS algorithm for any sample size, any number of categorical variables and any computational situations.

IS16 BAYESIAN ANALYSIS OF HIGH-DIMENSIONAL DATA
Session Organizer and Chair: Jaeyong Lee, Seoul National University, South Korea
Venue: LT52
Time: 17 Dec, 10:40-12:40

Comparison of Bayes shrinkage estimation methods for high-dimensional VAR models

Sung-Ho Kim
KAIST, South Korea

Abstract: Bayesian shrinkage estimation methods are useful for dealing with high-dimensional VAR models and they demonstrate better than or at least as good performances as such methods known in literature as Ridge and non-parametric methods among others. Some Bayesian methods using the MCMC process perform as well but at a considerable computation cost. In the talk, we will present new Bayesian estimation methods with conjugate and non-conjugate priors where the values of the shrinkage parameter are selected in such a way that the posterior mean of the log joint density of the data and the VAR coefficients is maximized. Through a simulation experiment under a wide scope of the conditions of the VAR model, we compared 10 estimation methods including our methods and investigated how their performances are affected by the model conditions such as model complexity, dimension, and the covariance structure of the noise variables.

A Bayesian test of independence for sparse contingency tables

Dalho Kim
Kyung- pook National University, South Korea

Abstract: We study the association between bone mineral density (BMD) and body mass index (BMI) when contingency tables are constructed from several U.S. counties, where BMD has three levels (normal, osteopenia and osteoporosis) and BMI has four levels (underweight, normal, overweight and obese). We use the Bayes factor (posterior odds divided by prior odds or equivalently the ratio of the marginal likelihoods) to construct the new test. Like the chi- squared test and Fisher's exact test, we have a direct Bayes test which is a standard test using data from each county. In our main contribution, for each county techniques of small area estimation are used to borrow strength across counties and via a hierarchical Bayesian model a pooled test of independence of BMD and BMI is obtained. Our pooled Bayes test is computed by performing a Monte Carlo integration using random samples rather than Gibbs samples. We have seen important differences among the pooled Bayes test, direct Bayes test and the Cressie-Read test which allows for some degree of sparseness, when the degree of evidence against independence is studied. As expected, we also found that the direct Bayes

test is sensitive to the prior specifications but the pooled Bayes test is not so sensitive. Moreover, the pooled Bayes test has competitive power properties.

A Bayesian parsimonious model search for high-dimensional variable selection

Chae Young Lim
Seoul National University, South Korea

Abstract: Bayesian variable selection models have been explored and developed as a powerful alternative to the classical approaches for model selection and comparisons using well known criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The Bayesian model search provides further insight into the uncertainty associated with the highly concerned features from model perspective such as the number of important variables, which is particularly informative for high-dimensional variable selection when p is quite large or exceeds N, the number of observations. In this work, we present a parsimonious model search via a combination of the L_0 and L_2 penalties, which is closely connected with the classical and most commonly used criteria for model selection and comparisons. We also consider an adaptive learning procedure for model search in large candidate sets. The method can accommodate missing data scenario, non-Gaussian responses, and presence of correlations between the observations.

Bayesian typhoon track prediction using wind vector data

Minkyu Han
Asan Medical Center, South Korea.

Abstract: In this paper, we predict the track of typhoons using Bayesian principal component regression model based on the wind field data. The data is obtained at each time point and we applied the Bayesian principal component regression model to conduct the track prediction based on the time point. We show that the results in this paper are close to the result of Japanese Meteorological Agency, it is possible to predict the track of typhoon very accurately using only statistical model.

IS17 YOUNG STATISTICIANS GROUP IN IASC (YSG-IASC)
Session Organizer and Chair: Han-Ming Wu, Tamkang University, Taiwan
Venue: LT51
Time: 17 Dec, 16:00-18:00

Error bounds for sequential Monte Carlo methods on multimodal distributions

Daniel Paulin
National University of Singapore, Singapore

Abstract: Multimodal distributions pose formidable challenges for Bayesian inference methods. In this talk, we present some new theory on the effectiveness Sequential Monte Carlo methods for such distributions. This is joint work with Ajay Jasra and Alexandre H. Thiery."

Sensible functional LDA

Ci-Ren Jiang
Institute of Statistical Science, Academia Sinica, Taiwan
E-mail: cirenjiang@stat.sinica.edu.tw

Keywords: Classification, Functional/Longitudinal Data, Linear Discriminant Analysis, Principal Component Analysis, Smoothing.

Abstract: The focus of this work is to extend Fisher's linear discriminant analysis (LDA) to both densely recorded functional data and sparsely observed longitudinal data. Despite the difference in sampling schemes, functional data and longitudinal data come from similar source. It is thus sensible to develop a unified approach for them. Not only is the noninvertibility issue of the covariance operator emerging from this extension managed, but two scenarios on the range spaces of within-subject covariance function and of between-class covariance function are investigated to seek the optimal LDA directions for subsequent classification. A conditional expectation technique is employed to tackle the challenge of projecting longitudinal data to LDA directions. Further, asymptotic properties of the proposed estimators are studied and the performance of this new approach is demonstrated with numerical examples. For high dimensional data, our numerical investigations have shown that our approach with less computational cost still yields comparable performance with linear support vector machines, especially when the sample size is moderately large. Therefore, the proposed approach seems quite competitive and promising in this era of big data.

On classification for functional data analysis

Diego Andres Perez Ruiz
School of Mathematics & University of Manchester, UK
E-mail: diego.perezruiz@manchester.ac.uk

Keywords: Functional Data, Supervised Classification, Functional Principal Components.

Abstract: A nonparametric tool for discriminating functional data based on the semi metric from the functional principal components is proposed. This problem is known as supervised classification applied to a sample of curve and there is a sizable literature on methods for classifying functional data. It is proposed to use this semi metric for supervised classification and construct a kernel-based classification. Simulation studies show that this procedure perform well and in some cases better than existing supervised classification methods for functions, resulting in high accuracy. An empirical example using real data is also provided to illustrate the results.

Analysis of cortical thickness with medical imaging data of very high and varying dimensions per subject

Elvan Ceyhan
Koç University, Turkey

Abstract: Neuropsychiatric disorders are manifested in anatomical shape differences in cortical structures. Labeled Cortical Distance Mapping (LCDM) is a powerful tool in quantifying such morphometric differences. LCDM characterizes the morphometry of the laminar cortical mantle of cortical structures. Specifically, LCDM data are distances of labeled gray matter (GM) voxels with respect to the gray/white matter cortical surface. Depending on the voxel resolution, LCDM distances can be very large and are of differing dimension per subject. Volumes and descriptive measures (such as means and variances for each subject) based on LCDM distances provide descriptive summary information on some of the shape characteristics. To use more of this information, we pool (merge) LCDM distances from subjects in the same group. These pooled distances can help detect morphometric differences between groups, but do not provide information about the locations of such differences in the tissue in question. We specify the types of alternatives for which the tests are more sensitive. We also show that the pooled LCDM distances provide powerful results for group differences in distribution of LCDM distances. We also censor LCDM distances at a fixed increment size; i.e., we keep distances less than the censoring distance at each increment step. The analysis of censored distances provides more information about the location of morphometric differences (left-right morphometric asymmetry and morphometric group differences) compared to pooled distances.

IS18 HIGH DIMENSION HYPOTHESIS TESTING, CHANGE POINT, VARIABLE SELECTION WITH CATEGORICAL VARIABLES
Session Organizer and Chair: Guangming Pan, Nanyang Technological University, Singapore
Venue: LT52
Time: 18 Dec, 15:50-17:50

Bias and variance reduction in estimating the proportion of true null hypotheses

Tong Tiejun
Hong Kong Baptist University, Hong Kong

Abstract: When testing a large number of hypotheses, estimating the proportion of true nulls, denoted by pi0, becomes increasingly important. This quantity has many applications in practice. For instance, a reliable estimate of pi0 can eliminate the conservative bias of the Benjamini–Hochberg procedure on controlling the false discovery rate. It is known that most methods in the literature for estimating pi0 are conservative. Recently, some attempts have been paid to reduce such estimation bias. Nevertheless, they are either over bias corrected or suffering from an unacceptably large estimation variance. In this paper, we propose a new method for estimating pi0 that aims to reduce the bias and variance of the estimation simultaneously. To achieve this, we first utilize the probability density functions of false-null p-values and then propose a novel algorithm to estimate the quantity of pi0. The statistical behavior of the proposed estimator is also investigated. Finally, we carry out extensive simulation studies and several real data analysis to evaluate the performance of the proposed estimator. Both simulated and real data demonstrate that the proposed method may improve the existing literature significantly.

Model selection with categorical predictors

Peng Heng
Hong Kong Baptist University, Hong Kong

Abstract: In this paper, we study the problem of model selection with category predictors.
The categories are firstly ordered by a clustering method based on the parameter estimators via least squares. Then through simple transformation, we change the problem of identifying categories combination to variable selection for linear regression model. Numerical algorithms are described in details with the tuning parameter λ selected by BIC. Theoretical results are established for the two cases where the observation numbers ni's are diverging or bounded. It is proved that the estimator possesses consistency and sparsity. We also point out the advantages of the estimates by our proposal compared to the least squares method. Finite sample performance

of the proposed approach is verified by Monte Carlo simulations. We also analyze a real data example to demonstrate the usefulness of the proposal.

Estimation of the change point in the high-dimensional covariance matrix

Yang Qing
Nanyang Technological University, Singapore

Abstract: The change point estimation problem has been widely considered for mean vectors, while covariance matrices are less studied, especially in the high-dimensional setting. We propose a new method to estimate the true change point in the high-dimensional covariance matrix and develop the estimator's asymptotic properties. Numerical studies support the effectiveness of our estimator.

Model-free feature screening for ultrahigh dimensional discriminant analysis

Wei Zhong
Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, Xiamen University, China
Email: wzhong@xmu.edu.cn

Abstract: This work is concerned with marginal sure independence feature screening for ultrahigh dimensional discriminant analysis. The response variable is categorical in discriminant analysis. This enables us to use conditional distribution function to construct a new index for feature screening. In this paper, we propose a marginal feature screening procedure based on empirical conditional distribution function. We establish the sure screening and ranking consistency properties for the proposed procedure without assuming any moment condition on the predictors. The proposed procedure enjoys several appealing merits. First, it is model-free in that its implementation does not require specification of a regression model. Second, it is robust to heavy-tailed distributions of predictors and the presence of potential outliers. Third, it allows the categorical response having a diverging number of classes in the order of O(n^{κ}) with some $\kappa \geq 0$. We assess the finite sample property of the proposed procedure by Monte Carlo simulation studies and numerical comparison. We further illustrate the proposed methodology by empirical analyses of two real-life data sets.
(Joint work with Hengjian Cui and Runze Li.)

IS19 DISTANCE AND SUBSPACE LEARNING, INTERACTION SCREENING
Session Organizer and Chair: Zheng Tracy Ke, University of Chicago, USA
Venue: LT52
Time: 19 Dec, 9:00-11:00

Convex regularization for low rank tensor estimation

Ming Yuan
University of Wisconsin-Madison, USA

Abstract: Many problems can be formulated as recovering a low-rank tensor. Although an increasingly common task, tensor recovery remains a challenging problem because of the delicacy associated with the decomposition of higher order tensors. We will introduce a general framework of convex regularization for low rank tensor estimation.

Learning subspaces of different dimensions

Lek-Heng Lim
University of Chicago, USA

Abstract: Modern data are often characterized by their principal subspaces and "subspace learning", i.e., statistical models for inferring mixtures of subspaces, has become a topic of interest. We will discuss some recent progress in learning (1) subspaces of different dimensions and (2) affine subspaces. The set of all subspaces of a given dimension is a well-known geometric object known as a Grassmannian. Our Bayesian model would depend on a clever embedding of Grassmannians of different dimensions into the unit sphere of relatively low dimension.
As will become apparent, such models invariably rest upon a notion of distance between subspaces. For two subspaces of the same dimension, there is a well-known intrinsic notion of distance -- the geodesic distance between two points on a Grassmannian. This is intrinsic in the sense that it does not depend on an embedding of the Grassmannain into some larger ambient space, and furthermore it can be related to principle angles and thus computed via the SVD. We will discuss intrinsic distances for (1) subspaces of different dimensions and (2) affine subspaces inspired respectively by algebraic geometry (Schubert varieties) and differential geometry (universal quotient bundle). Both are readily computable via SVD.

The first part of this talk is joint work with Lizhen Lin, Sayan Mukherjee, and Brian St. Thomas of Duke. The second part is joint work with Ke Ye of Chicago.

Innovated interaction screening for high-dimensional nonlinear classification

Zemin Zheng
University of Science and Technology in China, China

Abstract: This work is concerned with the problems of interaction screening and nonlinear classification in high-dimensional setting. We propose a two-step procedure, IIS-SQDA, where in the first step an innovated interaction screening (IIS) approach based on transforming the original p-dimensional feature vector is proposed, and in the second step a sparse quadratic discriminant analysis (SQDA) is proposed for further selecting important interactions and main effects and simultaneously conducting classification. Our IIS approach screens important interactions by examining only p features instead of all two-way interactions of order $O(p^2)$. Our theory shows that the proposed method enjoys sure screening property in interaction selection in the high-dimensional setting of p growing exponentially with the sample size. In the selection and classification step, we establish a sparse inequality on the estimated coefficient vector for QDA and prove that the classification error of our procedure can be upper-bounded by the oracle classification error plus some smaller order term. Extensive simulation studies and real data analysis show that our proposal compares favorably with existing methods in interaction selection and high-dimensional classification.
(Joint work with Yingying Fan, Yinfei Kong and Daoji Li)

Large covariance estimation for compositional data via composition-adjusted thresholding

Wei Lin
Peking University, China

Abstract: High-dimensional compositional data arise naturally in many application areas such as metagenomic data analysis. The observed data lie in a high-dimensional simplex and conventional statistical methods often fail to produce sensible results. In this article, we address the problem of high-dimensional compositional data analysis from a latent variable modeling perspective. We introduce a composition-adjusted thresholding (COAT) method for estimating the covariance structure of high-dimensional compositional data under the assumption that the basis covariance matrix is sparse. Our method is based on a decomposition relating the compositional covariance to the basis covariance, which is approximately identifiable as the dimensionality tends to infinity. The resulting procedure can be viewed as thresholding the sample centered log-ratio covariance matrix and hence is scalable for large covariance matrices. We rigorously characterize the identifiability of the covariance parameters, derive rates of convergence under the spectral norm, and provide theoretical guarantees on support recovery. Simulation studies demonstrate that the COAT estimator outperforms some naive thresholding estimators that ignore the unique features of compositional data. We apply the proposed method to the analysis of a microbiome data set

in order to understand the dependence structure of bacterial genera in the human gut.

IS20 NEW CHALLENGES IN FINANCIAL DATA
Session Organizer and Chair: Yingying Li, Hong Kong University of Science and Technology, Hong Kong
Venue: LT52
Time: 17 Dec, 16:00-18:00

The microstructural foundations of rough volatility models

Mathieu Rosenbaum
University Marie and Pierre Curie (Paris 6), France

Abstract: It has been recently shown that rough volatility models reproduce very well the statistical properties of low frequency financial data. In such models, the volatility process is driven by a fractional Brownian motion with Hurst parameter of order 0.1. The goal of this talk is to explain how such fractional dynamics can be obtained from the behaviour of market participants at the microstructural scales. Using limit theorems for Hawkes processes, we show that a rough volatility naturally arises in the presence of high frequency trading combined with metaorders splitting. This is joint work with Thibault Jaisson.

Solving the high-dimensional Markowitz optimization problem: when sparse regression meets random matrix theory

Yingying Li
Hong Kong University of Science and Technology, Hong Kong
E-mail: yyli@ust.hk

Abstract: To solve the high-dimensional Markowitz optimization problem, a new approach combining sparse regression and estimation of maximum expected return for a given risk level based on random matrix theory is proposed. We prove that under some sparsity assumptions on the underlying optimal portfolio, our estimated portfolio, the Response-estimated Sparse Regression Portfolio (ReSReP), asymptotically reaches the maximum expected return and meanwhile satisfies the risk constraint. To the best of our knowledge, this is the first time that these two goals are simultaneously achieved in the high-dimensional setting. The superior properties of ReSReP are demonstrated via simulation and extensive empirical studies. Joint work with Mengmeng Ao and Xinghua Zheng.

On the inference about the spectral distribution of high-dimensional covariance matrix based on noisy observations with applications to integrated covolatility matrix inference in the presence of microstructure noise

Xinghua Zheng
Hong Kong University of Science and Technology, Hong Kong

Abstract: In practice, observations are often contaminated by noise, making the resulting sample covariance matrix to be an information-plus-noise-type covariance matrix. Aiming to make inferences about the spectrum of the underlying true covariance matrix under such a situation, we establish an asymptotic relationship that describes how the limiting spectral distribution of (true) sample covariance matrices depends on that of information-plus-noise-type sample covariance matrices. As an application, we consider the inference about the spectrum of integrated covolatility (ICV) matrices of high-dimensional diffusion processes based on high-frequency data with microstructure noise. The (slightly modified) pre-averaging estimator is an information-plus-noise-type covariance matrix, and the aforementioned result, together with a (generalized) connection between the spectral distribution of true sample covariance matrices and that of the population covariance matrix, enables us to propose a two-step procedure to estimate the spectral distribution of ICV for a class of diffusion processes. An alternative estimator is further proposed, which possesses two desirable properties: it eliminates the impact of microstructure noise, and its limiting spectral distribution depends only on that of the ICV through the standard Marcenko-Pastur equation. Numerical studies demonstrate that our proposed methods can be used to estimate the spectrum of the underlying covariance matrix based on noisy observations.

To scale or not to scale? That is the question.

Bing-Yi Jing
Hong Kong University of Science and Technology, Hong Kong

Abstract: In regression problems, covariates may come in different types (e.g., numeric or categorical) and sizes (e.g., in meters or inches). In this talk, we shall discuss whether it is necessary to scale or standardize these covariates. If it is, how does one do this? Although different people have different suggestions, the issue of scaling has not been seriously studied in the literature. In this talk, we shall investigate this problem and report some findings.

IS21 APPLICATIONS OF MEMETIC AND METAHEURISTIC ALGORITHMS FOR SOLVING BIOMEDICAL PROBLEMS

Session Organizer and Chair: Weng Kee Wong, University of California Los Angeles, USA
Venue: Seminar Room 4
Time: 19 Dec, 9:00-11:00

Intelligent modelling for survival data with dependent censoring

Stefanie Biederrmann
University of Southampton, UK

Abstract: There are often reasons to believe that there may be dependence between the time to event and time to censoring in survival data, particularly in a medical context. The motivating example for this study relates to the survival of patients on the waiting list for a liver transplant where, broadly speaking, the most ill patients are prioritised to receive a transplant. Such censoring is known as dependent censoring and is in contrast to the assumption of non-informative censoring that underlies standard survival techniques. Because of identifiability issues, dependent censoring cannot be detected by statistical tests. A recent approach to assess the effect of dependent censoring is based on sensitivity analyses. We present some novel research in this area, including a new modelling approach and simulations.

Bayesian optimal design for ordinary differential equation models

Antony Overstall
University of Glasgow, UK

Abstract: Bayesian optimal design is considered for physical models derived from the (intractable) solution to a system of ordinary differential equations (ODEs).
Bayesian optimal design requires the minimisation of the expectation (over all unknown and unobserved quantities) of an appropriately chosen loss function. This can be non-trivial due to 1) the high dimensionality of the design space; and 2) the intractability of the expected loss. In this case, a further complication arises from the intractability of the solution to the system of ODEs.
We propose a strategy that employs a modification of the continuous coordinate exchange algorithm where a statistical emulator is employed to approximate the expected loss function, and a probabilistic solution to the system of ODEs. The strategy is demonstrated on several illustrative examples from the biological sciences.

Advancements in memetic computation with applications to real world applications

Yew Soon Ong
School of Computer Engineering, Nanyang Technological University, Singapore

Abstract: We are in an era where a plethora of computational problem-solving methodologies are being invented to tackle the diverse problems that are of interest to researchers. Some of these problems have emerged from real-life scenarios while some are theoretically motivated and created to stretch the bounds of current computational algorithms. Regardless, it is clear that in this new millennium a unifying concept to dissolve the barriers among these techniques will help to advance the course of algorithmic research. Interestingly, there is a parallel that can be drawn in memes from both socio-cultural and computational perspectives. The platform for memes in the former is the human minds while in the latter, the platform for memes is algorithms for problem-solving. In this context, memes can culminate into representations that enhance the problem-solving capability of algorithms. The phrase Memetic Computing has surfaced in recent years; emerging as a discipline of research that focuses on the use of memes
as units of information which is analogous to memes in a social and cultural context. One of the most popular early instantiations of Memetic Computation is memetic algorithms. In this talk, a comprehensive multi-facet survey and the roles of “meme” in computational intelligence is first reviewed. Subsequently, we take a peek into several state-of-the-art memetic algorithms and examined some recent frameworks and theoretic studies of memetic computation. Some successful applications of memetic computing methodologies for solving complex application in Biomolecular Systems are showcased.

A first-order algorithm for the A-optimal experimental design problem: a mathematical programming approach

Ahipasaoglu Selin Damla
Engineering System and Design, University of Technology and Design, Singapore

Abstract: We develop and analyze a first-order algorithm for the A-optimal experimental design problem. The problem is first presented as a special case of a parametric family of optimal design problems for which duality results and optimality conditions are given. Then, two first-order (Frank-Wolfe type) algorithms are presented, accompanied by a detailed time-complexity analysis of the algorithms and computational results on various sized problems.

IS22 ADVANCED STATISTICAL PARAMETRIC ANALYSIS
Session Organizer: Jialiang Li, National University of Singapore, Singapore
Session Chair: Binyan Jiang, Hong Kong Polytechnic University, Hong Kong
Venue: Seminar Room 1
Time: 18 Dec, 15:50-17:50

Limit laws for the Zagreb indices of several random graph models

Feng Qunqiang
University of Science and Technology of China, China

Abstract: A topological index is a map from the set of graphs to the set of real numbers. The Zagreb indices, which are a couple of well-known topological indices, were first introduced by chemists Gutman and Trinajstic (1972). In this talk, wet will show some limit laws for the Zagreb indices of three widely studied random graph models: random recursive trees, scale-free trees and classical Erdos-Renyi random graphs. The methods we use here include martingale limit theorems and Stein's method. This is joint work with Hu Zhishui and Su Chun.

Two stage multiple change-points detection in linear models

Jin Baisuo
University of Science and Technology of China, China

Abstract: A two-stage procedure for simultaneously detecting multiple change-points in linear regression is developed, using the model selection methods and a refining method. Consistency of the change-point estimates is established under mild conditions. The new procedure is fast and accurate as shown in simulation studies. Its applicability in two real situations were demonstrated via a well-log data and an ozone data.

Large dimensional integrated covariance matrix estimation and its application in portfolio choice

Liu Cheng
Wuhan University, China

Abstract: The estimator of integrated covariance matrix (ICM) plays a crucial role in Markowitz (1952) portfolio selection. In this projection, we propose a new estimator of high dimensional ICM by the random matrix theory for high dimensional portfolio selection based on high frequency data. We show that (1) high frequency data can improve the high dimensional portfolio allocation; (2) the portfolio based on our estimator yield a significantly lower volatility than others. (3) Our estimator for ICM is also suitable for the case that the

dimensional of assets is bigger than the sample size. The gain of our approach in portfolio choice are demonstrated numerically through simulation studies and real data analysis. We also compare the performance of high and low frequency data in portfolio choice problems.

Sequential change-point detection based on nearest neighbors

Hao Chen
University of California, Davis, USA

Abstract: As we observe the dynamics of social networks over time, how can we tell if a significant change happens? We propose a new framework for the detection of change-points as data are generated. The approach utilizes nearest neighbor information and can be applied to ongoing sequences of multivariate data or object data. Different stopping times are compared and one relies on recent observations is recommended. An accurate analytic approximation is obtained for the average run length when there is no change, facilitating its application to real problems.

IS23 SYSTEMIC RISK MODELING BASED ON HIGH-FREQUENCY DATA
Session Organizer and Chair: Sergey Ivliev, Perm State University, Russia
Venue: LT52
Time: 17 Dec, 13:40-15:40

Risk Related Brain Regions Detection and Individual Risk Classification with 3D Image FPCA

Ying Chen
National University of Singapore, Singapore

Abstract: Understanding how people make decisions among risky choices has attracted much attention of researchers in economics, psychology, and neuroscience. While economists try to evaluate individual's risk preference through mathematical modeling, neuroscientists answer the question by exploring the neural activities in brain. We propose a novel model-free method, 3-dimensional image functional principal component analysis (3DIF), to provide a connection between active risk related brain region detection and individual's risk preference. The 3DIF methodology is directly applicable to 3D image data without artificial vectorization or mapping and simultaneously guarantees the contiguity of risk related brain regions rather than discrete voxels. Simulation study evidences an accurate and reasonable region detection using the 3DIF method. In real data analysis, 5 important risk related brain regions are detected, including parietal cortex (PC), ventrolateral prefrontal cortex (VLPFC), lateral orbifrontal cortex (lOFC), anterior insula (aINS) and dorsolateral prefrontal cortex (DLPFC), while the alternative methods only identify limited risk related regions. Moreover, the 3DIF method

is useful for extraction of subjective specific signature scores that carry explanatory power for individual's risk attitude. In particular, the 3DIF method perfectly classifies both strongly and weakly risk averse subjects for in-sample analysis. In out-of-sample experiment, it achieves 73-88% overall accuracy, among which 90-100% strongly risk averse subjects and 49-71% for weakly risk averse subjects are correctly classified with leave-k-out cross validations. (This is a joint work with Wolfgang Karl Härdle, Qiang He and Piotr Majer.)

Systemic risk in dynamical networks with stochastic failure criterion

Ling Feng
A*STAR, Singapore

Abstract: Complex non-linear interactions between banks and assets we model by two time-dependent Erdős-Renyi network models where each node, representing a bank, can invest either to a single asset (model I) or multiple assets (model II). We use a dynamical network approach to evaluate the collective financial failure —systemic risk— quantified by the fraction of active nodes. The systemic risk can be calculated over any future time period, divided into sub-periods, where within each sub-period banks may contiguously fail due to links to either i) assets or ii) other banks, controlled by two parameters, probability of internal failure p and threshold Th ("solvency" parameter). The systemic risk decreases with the average network degree faster when all assets are equally distributed across banks than if assets are randomly distributed. The more inactive banks each bank can sustain (smaller Th), the smaller the systemic risk —for some Th values in I we report a discontinuity in systemic risk. When contiguous spreading becomes stochastic ii) controlled by probability p2 —a condition for the bank to be solvent (active) is stochastic— the systemic risk decreases with decreasing p2. We analyse the asset allocation for the U.S. banks.

Financial market instability: price shocks and the role of HFT

Sergey Ivliev
Department of Information System and Mathematical Methods in Economics, Perm State University, Perm, Bukireva 15, Russia
Email: ivliev@prognoz.ru

We present a study of stock market instability in a form of price shocks (jumps). In particular we want to answer two research questions (i) what are the characteristics of the shocks in Russian stock market prices? and (ii) what is the origin of these shocks with regard of the role of HFT?
We investigate ultra-high-frequency financial data to detect large but temporally localized price movements for 29 blue chip stocks traded at the Moscow Interbank Currency Exchange (MICEX) in a period of 82 trading days in 2010. We consider three different methods and time scales for identifying the shocks, namely one hour, one minute, and few ticks, and we call them macro, meso, and micro shocks. We find that the number of shocks scales as

a power law of the number of orders (or trades) of the stock. We show that, while one fifth of the micro shocks are embedded in a meso shock, only 1% of the meso shocks are embedded in a macro shocks, indicating that shocks on different scales might have different origin. We then study the dynamics of four key market metrics, namely price, buy/sell imbalance, trading volume and bid-ask spread, around large intraday price changes by considering separately macro and meso shocks.
Secondly, we present preliminary results of the HFT behavior analysis at major blue chip stock in four minute interval around identified 8-sigma meso shocks. We found that HFTs strongly modify their trading and order placement strategy during those shocks, increase trading, order placement and cancellations, and widen spreads. On a one minute time scale, no sign of precursors of shocks based on HFTs switching strategy was identified. Obtained results could usefully shed some light on the controversial nature of price jumps.
This talk represents work with F. Lillo, M. Frolova

IS24 FUNCTIONAL DATA ANALYSIS AND ITS APPLICATIONS
Session Organizer and Chair: Ci-Ren Jiang, Academia Sinica, Taiwan
Venue: LT50
Time: 18 Dec, 10:20-12:20

Regularized principal component analysis for spatial data

Hsin-Cheng Huang
Institute of Statistical Science, Academia Sinica, Taiwan
Email: hchuang@stat.sinica.edu.tw

Keywords: Alternating direction method of multipliers, empirical orthogonal functions, fixed rank kriging, Lasso, non-stationary spatial covariance estimation, orthogonal constraint, smoothing splines.

Abstract: In many atmospheric and earth sciences, it is of interest to identify dominant spatial patterns of variation based on data observed at p locations with n repeated measurements. While principal component analysis (PCA) is commonly applied to find the patterns, the eigenimages produced from PCA may be noisy or exhibit patterns that are not physically meaningful when p is large relative to n. To obtain more precise estimates of eigenimages (eigenfunctions), we propose a regularization approach incorporating smoothness and sparseness of eigenfunctions, while accounting for their orthogonality. Our method allows data taken at irregularly spaced or sparse locations. In addition, the resulting optimization problem can be solved using the alternating direction method of multipliers, which is computationally fast, easy to implement, and applicable to a large spatial dataset. Furthermore, the estimated eigenfunctions provide a natural basis for representing the underlying spatial process in a spatial random-effects model, from which spatial covariance function estimation and spatial prediction can be efficiently

performed using a regularized fixed-rank kriging method. Finally, the effectiveness of the proposed method is demonstrated by several numerical examples.

Multi-dimensional functional principal component analysis

Chen Lu-Hung
National Chung Hsing University, Taiwan

Abstract: Functional principal component analysis (FPCA) is one the most commonly employed approaches in functional/longitudinal data analysis and we extend it to conduct d-dimensional functional/longitudinal data analysis. The local linear smoothing technique is employed to perform estimation because of its capabilities of handling data with different sampling schemes and of performing large-scale smoothing in addition to its nice theoretical properties. Several computational techniques such as the modern GPGPU (general-purpose computing on graphics processing units) architecture are applied to perform parallel computation to save computation time and resources. We also show that the proposed estimators can achieve the classical nonparametric rates for longitudinal data; the optimal convergence rates can be achieved if the number of observations per sample function is of the order (n/logn)d/4 for functional data. The performance of our approach is further demonstrated with a simulation and two real data examples.

Shrinkage estimation for multilevel functional data: decoding resting state functional connectivity

Haochang Shou
Department of Biostatistics and Epidemiology, University of Pennsylvania, USA
Email: hshou@upenn.edu

Keywords: Resting-state fMRI, Shrinkage estimator, Connectivity, Parcellation, Mutlilevel functional model, Measurement error correction;

Abstract: Resting-state functional magnetic resonance imaging (rs-fMRI) has been used to investigate synchronous activations in spatially distinct regions of the brain, which are thought to reflect functional systems supporting cognitive processes. However, imaging data are observed with measurement errors that come from both voxel-wise random noises and systematic errors that are spatially correlated. Simply taking average over the repeated scans is not enough to reduce the effect of systematic errors. We propose the imaging intra-class correlation (I2C2) coefficient to quantify reproducibility for various modalities of scan-rescan imaging data and assess the signal-to-noise ratios. We then use a set of empirical Bayes shrinkage estimators that dramatically improve subject-specific prediction by borrowing strength from the population information based on all other images in the study, including both voxel-wise and global estimators where the spatial correlation are accounted. We apply

the methods to predict the seed-based connectivity maps and perform subject-level brain parcellation using a publicly available scan-rescan dataset from 20 subjects. In either application, the shrinkage estimates increase the reliability and validity by up to 30% than those produced from raw estimates. The proposed methods can be used as a pre-processing step for further imaging analysis.

Simultaneous confidence bands for functional regression models

Chung Chang
National Sun Yat-sen University, Taiwan
Email: binchung10000@gmail.com

Abstract: In recent years, the field of functional data analysis (FDA) has attracted a great deal of attention and many interesting theories and applications have been reported. Within this field, how to estimate simultaneous confidence bands (SCB) for an unknown function has been the object of extensive study. Estimating SCB for the mean function has drawn lots of attention in the literature. In this talk, we will estimate SCB for the coefficient function in a more general function-on-scalar regression model (allowing covariates). We will propose a new wild bootstrap method to estimate SCB and the advantage of this method is that it can deal with the heterogeneity problem. We will provide theoretical justification, including sufficient conditions to ensure the asymptotic theorem, for the mean function case and simulation results for more general situations. We will also compare our bootstrap method with the traditional one.

IS25 MODEL SPECIFICATION AND SELECTION
Session Organizer: I-Ping Tu, Academia Sinica, Taiwan
Session chair: Su-Yun Huang, Academia Sinica, Taiwan
Venue: LT50
Time: 19 Dec, 9:00-11:00

A validated information criterion (VIC) to find the structural dimension

Yanyuan Ma
University of South Carolina, USA
Email: yanyuanma@stat.sc.edu

Abstract: A crucial component in performing sufficient dimension reduction is to determine the structural dimension of the reduction model. We propose a novel information criterion-based method to achieve this purpose, whose special feature is that when examining the goodness-of-fit of the current model, we need to obtain model evaluation by using an enlarged candidate model. Although the procedure does not require estimation under the enlarged model with dimension k + 1, the decision on how well the current

model with dimension k fits relies on the validation provided by the enlarged model. This leads to the name validated information criterion, calculated as VIC(k). The method is different from existing information criteria based model selection methods. It breaks free from the dependence on the connection between dimension reduction models and their corresponding matrix eigen-structures, which heavily relies on a linearity condition that we no longer assume. Its consistency is proved and its finite sample performance is demonstrated numerically. (Joint work with Xinyu Zhang.)

Extension of AUC for classification considering heterogeneity in distributions

Osamu Komori
The Institute of Statistical Mathematics, Japan

Keywords Area under the ROC curve; asymptotic normality and variance; Binary classification

Abstract: In the two-group classification problems, heterogeneity of probability distributions is a key factor to be paid attention to in order to improve the classification accuracy. In the previous studies, we considered a t-statistic-based approach based on a linear predictor, assuming that homogeneity for one group and heterogeneity for another group, which is often the case when we consider case-control studies. This time we propose a new statistical method based on AUC that allows for heterogeneity for the both two groups, which is often the case when dealing with various cancer datasets. We investigate the asymptotic variance of the proposed linear predictor and compare the classification accuracy with other existing methods. The performance is illustrated by several simulation studies based on small sample and high dimensional datasets.

Spontaneous learning for clustered data via gamma - power divergence

Shinto Eguchi
The Institute of Statistical Mathematics, Japan

Abstract: The clustering algorithm applying minimum divergence method is discussed to apply to high dimensional data. In particular we employ the gamma - power divergence for an exponential family that performs a spontaneous learning for such data sets. We observe that the minimization of the gamma - power divergence leads to local minima which properly correspond to the centers of clustered subsets. The statistical property for such spontaneous learning is explored from information geometric viewpoint.

Estimating high-dimensional multi-layered networks through penalized maximum likelihood

George Michailidis
University of Florida, USA

Abstract: Gaussian graphical models represent a good tool for capturing interactions between nodes represent the underlying random variables. However, in many applications in biology one is interested in modeling associations both between, as well as within molecular compartments (e.g. interactions between genes and proteins/metabolites). To this end, inferring multi-layered network structures from high-dimensional data provides insight into understanding the conditional relationships among nodes within layers, after adjusting for and quantifying the effects of nodes from other layers. We propose an integrated algorithmic approach for estimating multi-layered networks that incorporates a screening step for significant variables, an optimization algorithm for estimating the key model parameters and a stability selection step for selecting the most stable effects. The proposed methodology offers an efficient way of estimating the edges within and across layers iteratively, by solving an optimization problem constructed based on penalized maximum likelihood (under a Gaussian assumption). The optimization is solved on a reduced parameter space that is identified through screening, which remedies the instability in high-dimension. Theoretical properties are considered to ensure identifiability and consistent estimation of the parameters and convergence of the optimization algorithm, despite the lack of global convexity. The performance of the methodology is illustrated on synthetic data sets and on an application on gene and metabolic expression data for patients with renal disease.

IS26 FRONTIERS IN STATISTICAL GENOMICS
Session Organizer and Chair: Hsin-Chou Yang, Academia Sinica, Taiwan
Venue: Stephen Riady Global Learning Room
Time: 17 Dec, 16:00-18:00

Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants

Sungho Won
Department of Public Health Science, Seoul National University, Seoul, Korea
Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea

Abstract: Background: Many disease phenotypes are outcomes of the complicated interplay between multiple genes, and multiple phenotypes are affected by a single or multiple genotypes. Therefore, joint analysis of multiple phenotypes and multiple markers has been considered as an efficient strategy for genome-wide association analysis, and in this work we propose an

omnibus family-based association test for the joint analysis of multiple genotypes and multiple phenotypes.
Results: The proposed test can be applied for both quantitative and dichotomous phenotypes, and it is robust under the presence of population substructure, as long as large-scale genomic data is available. Using simulated data, we showed that our method is statistically more efficient than the existing methods, and the practical relevance is illustrated by application of the approach to obesity-related phenotypes.
Conclusions: The proposed method may be more statistically efficient than the existing methods. The application was developed in C++ and is available at the following URL: http://healthstats.snu.ac.kr/~swon/software/software.php.

A fast algorithm for DNA methylation calling based on binomial mixtures

Agus Salim
La Trobe University, Melbourne, Australia

Abstract: Bissulte sequencing technology has allowed genome-wide interrogations of methylation profiles. While the technology is quite advance, it is still not perfect. Problems common to this kind of data include sequencing errors, incomplete cytosine conversion and batch effects. We develop a fast methylation calling algorithm for Bisulfite sequencing data. The software uses local likelihood smoothing to the estimate the methylation levels, with binomial mixtures to model the incomplete cytosine conversion. Novel features of the software include capacity to correct for incomplete conversion of unmethylated cytosines, adjustment for non-uniform reads quality and removal of non-experimental systematic variations. To demonstrate the algorithm, we use two replicates of whole genome shotgun bisulfite sequencing (WGBS) of H1 cell lines (GSE16256). We compare our algorithm to several existing algorithms and demonstrated that our algorithm has better consistency when making calls within biological replicates.

A fast and effective W-test for SNP-SNP interaction search in GWAS with application on Bipolar disorder

Maggie Haitian Wang
The Chinese University of Hong Kong, Hong Kong

Abstract: Genetic association study has the objective of identification of disease susceptible loci from genome-wide data. In recent years, the challenge remains in the revealing of the epistasis effect that underlies complex diseases. Many interaction methods were suggested, seeking for a balance among the complexity of methods, efficiency of computation, interpretability of results and feasibility of application for the ever increasing data volume. In this talk, we will introduce a fast and effective W-test, which has an odds ratio interpretation comparing the distributional difference between cases and controls. We will demonstrate its superior performance among alternative methods in simulated data sets under different genetic

architectures and varying sample size. The W-test is applied on two real GWAS bipolar disorder data sets using a three-step procedure. Besides successfully replicating the main effect previously reported, the W-test is able to identify interesting gene-gene interactions that can be validated using two independent GWAS data sets. The validated gene-gene interaction pairs reside in key neuro-function pathways, which cannot be found from main effect. They contribute to the completion of the genetic heritability picture of bipolar disorder, and provide valuable pharmaceutical targets for treatment of the disease.

Exon-seq analysis reveals adverse effect of targeting drug treatment associated genomic variations in lung adenocarcinoma

Ya-Hsuan Chang
Kim Forest Enterprise, 4F, 128, Xinhu 2nd Road, Taipei, Taiwan

Keywords: EGFR, tyrosine kinase inhibitors, adverse effects, single nucleotide variants.

Abstract: EGFR tyrosine kinase inhibitors (EGFR-TKIs) are the effective targeting therapies for lung cancer patients. TKI treatments improved progression-free survival of patients with EGFR activating mutations. However, some patient had adverse effects such as hepatotoxicity and rejected TKI treatments. Whole exon sequencing approach provided variants information of genes and was a powerful tool in genomic researches. In this study, exon-seq analysis was used in 10 patients with hepatotoxicity and 48 patients without hepatotoxicity after TKI treatment. Results showed that 1,378 nonsynonymous single nucleotide variants (SNVs) had significant different frequencies between hepatotoxicity and non-hepatotoxicity groups. In addition, 3 SNVs were located in genes involved in the metabolism pathway of EGFR-TKIs. Therefore, two SNVs (one is novel and another is known) were more susceptible for attributing to hepatotoxicity. Findings of this study may provide the reference for treatment selection after validation in a prospective cohort.

IS27 STATISTICAL METHODS FOR INTRACTABLE LIKELIHOOD FUNCTIONS

Session Organizer and Chair: Scott Sisson, University of New South Wales, Australia
Venue: Seminar Room 1
Time: 17 Dec, 10:40-12:40

Bayesian inference and model choice for low count time series models with intractable likelihoods

Chris Drovandi
Queensland University of Technology, Australia

Abstract: Here new pseudo-marginal algorithms are presented to perform Bayesian inference and model choice for low count time series models with intractable likelihoods. These methods involve including the alive particle filter within Markov chain Monte Carlo and sequential Monte Carlo algorithms. The particle filter allows for sequential matching of simulated with observed data one-at-a-time, resulting in closer matching compared with standard approximate Bayesian computation. Unlike some competing approaches, our methods are able to accommodate partially observed data and a certain class of non-Markovian models. This is joint work with Tony Pettitt and Roy McCutchan.

Exact ABC using importance sampling

Robert Kohn
University of New South Wales, Australia

Keywords: Approximate Bayesian Computation, Debiasing, Ising model, Marginal likelihood Estimate, Unbiased likelihood Estimate

Abstract: Approximate Bayesian Computation (ABC) is a powerful method for carrying out Bayesian inference when the likelihood is computationally intractable.
However, a drawback of ABC is that it is an approximate method that induces a systematic error because it is necessary to set a tolerance level to make the computation tractable. The issue of how to optimally set this tolerance level has been the subject of extensive research.
This paper proposes an ABC algorithm based on importance sampling that estimates expectations with respect to the exact posterior distribution given the observed summary statistics. This overcomes the need to select the tolerance level. By exact we mean that there is no systematic error and the Monte Carlo error can be made arbitrarily small by increasing the number of importance samples. We provide a formal justification for the method and study its convergence properties.
The method is illustrated in two applications and the empirical results suggest that the proposed ABC based estimators consistently converge to the true values as the number of importance samples increases.
Our proposed approach can be applied more generally to any importance sampling problem where an unbiased estimate of the likelihood is required.
Joint work with Minh Ngoc Tran

Variational Bayes with intractable likelihood

Minh-Ngoc Tran
University of Sydney, Australia

Abstract: Variational Bayes (VB) is rapidly becoming a popular tool for Bayesian inference in statistical modeling. However, the existing VB

algorithms are restricted to cases where the likelihood is tractable, which precludes the use of VB in many interesting models such as in state space models and in approximate Bayesian computation (ABC), where application of VB methods was previously impossible. This paper extends the scope of application of VB to cases where the likelihood is intractable, but can be estimated unbiasedly. The proposed VB method therefore makes it possible to carry out Bayesian inference in many statistical models, including state space models and ABC. The method is generic in the sense that it can be applied to almost all statistical models without requiring a model-based derivation, which is a drawback of many existing VB algorithms. This is joint work with David Nott and Robert Kohn.

IS28 STATISTICAL METHODS FOR INTRACTABLE LIKELIHOOD FUNCTIONS

Session Organizer and Chair: Robert Kohn, University of New South Wales, Australia
Venue: LT51
Time: 18 Dec, 10:20-12:20

Likelihood-based inference for symbolic data

Scott A. Sisson
University of New South Wales, Australia

Abstract: Computational challenges arise when performing likelihood-based analyses of very large datasets. Symbolic data analysis is one approach to circumvent this problem which partitions the original dataset into a moderate number of symbols, each in some distributional form. These symbols are then analysed directly as ``data'' where each datapoint additional now has internal variation. This talk will present a new mixture-model based likelihood construction for the analysis of interval-valued symbolic data that, unlike existing methods, accounts for the generative construction of each symbol from the underlying dataset. The performance of this approach is assessed through a simulated and real data analysis of credit card debt.

Bayesian inference in nonlinear structural econometric models with intractable likelihoods

Marcel Scharth
University of Sydney, Australia

Abstract: We introduce a new Markov chain Monte Carlo (MCMC) sampler called the Markov Interacting Importance Sampler (MIIS). The MIIS sampler uses conditional importance sampling (IS) approximations to jointly sample the current state of the Markov Chain and estimate conditional expectations, possibly by incorporating a full range of variance reduction techniques. We

compute Rao-Blackwellized estimates based on the conditional expectations to construct control variates for estimating expectations under the target distribution. The control variates are particularly efficient when there are substantial correlations between the variables in the target distribution, a challenging setting for MCMC. An important motivating application of MIIS occurs when the exact Gibbs sampler is not available because it is infeasible to directly simulate from the conditional distributions. In this case the MIIS method can be more efficient than a Metropolis-within-Gibbs approach. We also introduce the MIIS random walk algorithm, designed to accelerate convergence and improve upon the computational efficiency of standard random walk samplers. Simulated and empirical illustrations for Bayesian analysis show that the method significantly reduces the variance of Monte Carlo estimates compared to standard MCMC approaches, at equivalent implementation and computational effort.

A flexible spectral approach to nonparametric estimation of spatial data

Sally Wood
University of Sydney, Australia

Abstract: Likelihoods of flexible models for phenomenon such as the spread of disease across space and time are complex to evaluate and sometimes intractable. In this paper we develop a novel non-parametric spatial temporal model for data generated by possibly non-stationary processes. The non-stationarity may be with respect to time and space or it may be the product of other covariates. This research is motivated by the need to identify and predict influenza “signatures” across Australia. The temporal modelling is done in the spectral domain while a non-stationary Gaussian process prior is placed across the spatial surface. These temporal and spatial aspects are combined in a mixture formulation. The components in the mixture are the time-varying spectra at a particular location while mixing weights are parameterised to depend upon space and as well as other conditions such as weather or demographic data. Such a formulation is scales well to a high dimensional covariate space. The Whittle likelihood is used as an approximation to the true likelihood. The frequentist properties of the technique are examined via simulation. Initial analysis of the real example reveals that there are distinct influenza signatures in Australia and that these signatures depend not only on space and time but also on covariates such as workforce movement patterns.

Delayed-Acceptance strategies for speeding up Pseudo-Marginal MCMC

Alex Thiery
National University of Singapore, Singapore

Abstract: The Pseudo-marginal Metropolis-Hastings algorithm can be used for exploring a Bayesian posterior density when only an unbiased estimate of the likelihood is available. Delayed-Acceptance strategies can then be applied

when it is computationally expensive to calculate this unbiased stochastic estimate but a computationally cheap deterministic approximation is available. We provide a framework for incorporating relatively general deterministic approximations into the theoretical analysis of high-dimensional targets. The results provide insight into the effect of the accuracy of the deterministic approximation and the nature of the stochastic approximation on the efficiency of the delayed acceptance algorithm. The theory also informs a practical strategy for algorithm tuning.

IS29 THEORETICAL FOUNDATION OF BIG DATA
Session Organize: Guang Cheng, Purdue University, USA
Session Chair: Jialiang Li, National University of Singapore
Venue: Stephen Riady Global Learning Room
Time: 18 Dec, 13:30-15:30

Screening interactions for ultra-dimensional data

Hao Helen Zhang
University of Arizona

Abstract: In high dimensional regressions, the problem of interaction selection poses challenges both in computation and theory. We investigate some key issues and reveal new interesting theoretical results. A new class of methods are proposed for interaction screening, which are feasible for high dimensional data even when the dimension increases exponentially fast with the sample size. Numerical examples are shown to demonstrate promising performance of the new methods.

Bayes theory and methods for large networks

Debdeep Pati
Florida State University, USA

Abstract: Data available in the form of massive networks are increasingly becoming common in modern applications ranging from brain remote activity, protein interactions, web applications, social networks to name a few. Estimating large networks calls for structured dimension reduction and estimation in stylized domains, necessitating new tools for model based inference and theory. In this talk, we develop efficient computational approaches for estimating the parameters of a stochastic block model from a Bayesian perspective, when the number of communities are unknown. We also undertake a theoretical investigation of the posterior distribution of the parameters and show consistent detection of the underlying communities. En route, we develop geometric embedding techniques to exploit the lower dimensional structure of the parameter space which may be of independent interest. The methods are illustrated on simulated and real data examples.

Divide and conquer Gaussian process regression

Anirban Bhattacharya
Texas A&M University, USA

Abstract: Gaussian process regression is popularly used as a flexible function estimation procedure in Bayesian nonparametrics. However, posterior sampling becomes increasingly difficult with increasing number of observations due to the need to invert a large kernel matrix. A variety of approximation schemes has been proposed to obviate this issue, though theoretical guarantees are not available. We study theoretical properties of the ``divide and conquer'' procedure, where the data is split into subsamples and the posteriors across the subsamples are aggregated to obtain an aggregated posterior. We investigate ``closeness'' of the aggregated posterior to the true posterior in Wasserstein distance and provide optimality conditions on the number of subsamples.

On the statistical analysis of first-order optimization methods

Yiyuan She
Florida State University, USA

Abstract: Due to the explosion of large-scale datasets in modern statistical applications, people are often in favor of first-order optimization methods to obtain an estimator for complex learning tasks. This talk presents non-asymptotic statistical analysis of a class of regularized estimators defined in this way. The associate problems are not necessarily convex, and the estimators do not necessarily guarantee functional local optimality. We are able to show sharp oracle inequalities under a new type of less-demanding regularity conditions. The sequence of iterates can decay to the desired statistical accuracy geometrically fast. A progressive regularization technique proves to be useful to relax the conditions and results in a scalable screening method. Our results reveal different benefits brought by convex and nonconvex types of shrinkage.

IS30 NATURE-INSPIRED META-HEURISTIC APPROACHES AND THEIR APPLICATIONS IN DESIGNS OF EXPERIMENTS

Session Organizer and Chair: Frederick Kin Hing Phoa, Academia Sinica, Taiwan
Venue: Seminar Room 2
Time: 18 Dec, 15:50-17:50

Nature-inspired meta-heuristic algorithms for generating optimal experimental designs

Weng Kee Wong
Department of Biostatistics, UCLA, USA

Keywords: approximate design, exact design, equivalence theorem, information matrix, multipleobjective optimal design

Abstract: Nature-inspired meta-heuristic algorithms are increasingly studied and used in many disciplines to solve high-dimensional complex optimization problems in the real world. It appears relatively few of these algorithms are used in mainstream statistics even though they are simple to implement, very flexible and able to find an optimal or a nearly optimal solution quickly. Frequently, these methods do not require any assumption on the function to be optimized and the user only needs to input a few tuning parameters I will demonstrate the usefulness of some of these algorithms for finding different types of optimal designs for nonlinear models in dose response studies. Algorithms that I plan to discuss are more recent ones such as Cuckoo and Particle Swarm Optimization. I also compare their performances and advantages relative to deterministic state-of-the art algorithms.

The swarm intelligence based (SIB) method and its applications to statistics

Frederick Kin Hing Phoa
Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Natural heuristic methods, like the particle swarm optimization and many others, enjoy fast convergence towards optimal solution via a series of inter-particle communication. Such methods are common for the optimization problem in engineering, but few in statistics problem. It is especially difficult to implement in some fields of statistics as the search spaces are mostly discrete, while most natural heuristic methods require continuous search domains. This talk introduces a new method called the Swarm Intelligence Based (SIB) method for optimization in statistics problems, featuring the searches within discrete space. Such fields include experimental designs, community detection, change-point analysis, variable selection, etc. The SIB method is a natural heuristic method that includes the MIX and MOVE operations, which combines target units and selects the best units respectively. This method is advantageous over the traditional particle swarm

optimization and many other heuristic approaches in the sense that it is ready for the search of both continuous and discrete domains, and its global best particle is guaranteed to monotonically move towards the optimum.

A chemical reaction-inspired optimization algorithm for biomedical sciences

Albert Y.S. Lam
The University of Hong Kong, Hong Kong
E-mail: albertlam@ieee.org

Keywords: Chemical reaction optimization, metaheuristic, nature-inspired algorithm.

Abstract: We encounter optimization problems in our daily lives and in research in various fields. Some of them are so hard that we can, at best, approximate the best solutions with (meta-)heuristic methods. However, the huge number of optimization problems and the small number of generally acknowledged methods mean that more metaheuristics are needed to fill the gap. A new metaheuristic, called Chemical Reaction Optimization (CRO), was proposed to solve these hard problems. It mimics the interactions of molecules in a chemical reaction to reach a low energy stable state. CRO has demonstrated its competitive edge over existing methods in solving many real-world problems. It has been successfully applied to problems in many disciplines, e.g., networking, communications, operations research, computing, finance, energy and environment, computational intelligence, etc. Therefore, it provides a new approach for solving optimization problems, especially those which may not be solvable with the few generally acknowledged approaches. In this talk, we focus on the applications of CRO to biomedical sciences.

IS31 CIRCULANT ORTHOGONAL ARRAYS: STRUCTURE AND APPLICATIONS TO FMRI EXPERIMENTS

Session Organizer and Chair: Yuan-Lung Lin, Academia Sinica, Taiwan
Venue: LT52
Time: 18 Dec, 10:20-12:20

Optimal experimental designs for fMRI via circulant biased weighing designs

Ching-Shui Cheng
Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Functional magnetic resonance imaging (fMRI) technology is popularly used in many fields for studying how the brain reacts to mental stimuli. The identification of optimal fMRI experimental designs is crucial for rendering precise statistical inference on brain functions. We develop a

general theory to guide the selection of fMRI designs for estimating a hemodynamic response function (HRF) that models the effect over time of the mental stimulus, and for studying contrasts between HRFs. We provide a useful connection between fMRI designs and circulant biased weighing designs, establish the statistical optimality of some well-known fMRI designs, and identify several new classes of fMRI designs. Construction methods of high-quality fMRI designs are also given. This is a joint work with Ming-Hung Kao.

Experimental designs for functional MRI with uncertain model matrix

Ming-Hung (Jason) Kao, Lin Zhou
School of Mathematical and Statistical Sciences, Arizona State University, USA

Abstract: Functional magnetic resonance imaging (fMRI) is a widely used technology for acquiring better knowledge on the inner workings of the human brain. The success of an fMRI study hinges on the quality of the selected experimental design. However, the identification and construction of high-quality fMRI designs are almost always arduous, and require much research. Here, we consider a modern fMRI experimental setting where the model matrix of the statistical model depends on the subject's probabilistic behavior during the experiment and is thus uncertain at the design stage. We propose an efficient approach to obtain good designs for this complex setting. Our approach consists of a design selection criterion, that is easy to evaluate, and a very efficient computer search algorithm. Through case studies, we show that our approach significantly outperforms a recently proposed method in terms of the computing time and the achieved design efficiency.

Circulant (almost-)orthogonal array

Yuan-Lung Lin
Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Orthogonal arrays have been widely used in designing experiments, but they do not exist for all users' required dimensions. Recently, circulant orthogonal arrays (COA) were proposed and applied in many fields such as stream cypher cryptanalysis. Since circulant Hadamard matrices, which can be viewed as two-level orthogonal arrays of strength two, have been conjectured nonexistence, circulant almost orthogonal arrays are considered in fMRI experiments. In this talk, we propose a series of new designs called circulant almost orthogonal arrays (CAOA). Complete difference system (CDS) is also introduced and applied for the construction of COA and CAOA. We not only prove the equivalence relation between CDS and COA, but also construct CAOA of any prime power symbols.

IS32 STATISTICAL METHODS AND APPLICATIONS
Session Organizer and Chair: Shaoli Wang, Shanghai University of Finance and Economics, China
Venue: Seminar Room 3
Time: 18 Dec, 15:50-17:50

Robust mixture regression and outlier detection via penalized likelihood

Weixin Yao,
University of California, Riverside, USA
Email: weixin.yao@ucr.edu

Abstract: Finite mixture regression models have been widely used for modelling mixed regression relationships arising from a clustered and thus heterogeneous population. The classical normal mixture model, despite of its simplicity and wide applicability, may fail dramatically in the presence of severe outliers. We pro- pose a robust mixture regression approach based on a sparse, case-specific, and scale-dependent mean-shift parameterization, for simultaneously conducting outlier detection and robust parameter estimation. A penalized likelihood approach is adopted to induce sparsity among the mean-shift parameters so that the outliers are distinguished from the good observations, and a thresholding-embedded Expectation-Maximization (EM) algorithm is developed to enable stable and efficient computation. The proposed penalized estimation approach is shown to have strong connections with other robust methods including the trimmed likelihood and the M-estimation methods. Comparing with several existing methods, the proposed methods show outstanding performance in numerical studies.

Clustering analysis of sparse categorical data with weighted similarity

Peng Zhang
Zhejiang University, China
Email: pengz@zju.edu.cn

Abstract: This talk discusses sparse categorical data clustering analysis. Three types of weighted similarity are defined based on which several clustering algorithms are investigated, such as similarity vector method, K-modes method and hierarchical clustering method. In the similarity vector method, cluster cores are formed from a starting point with the largest χ^2 statistic comparing observed and the expected similarity vectors. Clusters are allowed to contain common elements so that a data point may belong to different clusters. Attributes of which non-zeros contribute more to the similarity among the cluster than zeros are claimed the defining features of a cluster. The clustering structure and their characteristics are discovered in this approach. Two real data are presented and analyzed using the proposed algorithms and other methods for comparison. Concluding remarks and a plan for future work is given at the end.

Ensemble sufficient dimension folding methods on analyzing matrix-valued data

Yuan Xue
University of International Business and Economics, China
Email: yuanxue@uibe.edu.cn

Abstract: In this paper, we construct novel sufficient dimension folding methods on analyzing matrix formed data. Different from conventional vector-valued predictor, the predictor is a random matrix and contains many more random variables. Traditional dimension reduction methods fail to preserve the matrix structure of the reduced predictors. Dimension folding methods for matrix- and array-valued predictors can preserve the data structure and enhance the accuracy of the reduced predictor. We introduce folded-OPG ensemble estimator and two refined estimators, folded-MAVE ensemble and folded-SR ensemble. The folded-SR ensemble method mitigates the problem of deciding the number of slices. A modified cross validation method is used to determine the dimensions of CDFS. Simulated examples demonstrate the performances of the folded ensemble methods by comparing with existing inverse folded methods.

IS33 CORRELATED DATA AND FINANCIAL TIME SERIES

Session Organizer and Chair: Zhen Pang, The Hong Kong Polytechnic University, Hong Kong
Venue: LT51
Time: 19 Dec, 9:00-11:00

Modeling GARCH processes by the conditional quantile estimation

Guodong Li
Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong

Keywords: Bootstrapping approximation; diagnostic checking; GARCH process; Quantile regression.

Abstract: This paper proposes a novel transformation to change the conditional quantile estimation of GARCH processes into that of linear GARCH processes, and a numerically feasible estimating procedure is then obtained. In order to check the adequacy of fitted models, we adapt the signs and absolute values of residuals to construct three diagnostic tools. A bootstrapping approximation is also considered to approximate the complicated asymptotic distributions, and they are the first bootstrap-based portmanteau tests in the literature. Simulation experiments are conducted to illustrate the efficacy of the proposed conditional quantile estimation and

goodness-of-fit tests. A real example is applied to demonstrate the usefulness of our methodology.

Fence methods to genetic application

Thuan Nguyen
Oregon Health&Science University, USA

Abstract: Model search strategies play an important role in finding simultaneous susceptibility genes that are associated with a trait. More particularly, model selection via the information criteria, such as the BIC with modifications, has received considerable attention in quantitative trait loci (QTL) mapping. However, such modifications often depend upon several factors, such as sample size, prior distribution, and the type of experiment, e.g., backcross, intercross. These changes make it difficult to generalize the methods to all cases. The fence method avoids such limitations with a unified approach, and hence can be used more broadly. In this talk, the method is studied in the case of backcross experiments (BE). In particular, a variation of the fence, called restricted fence (RF), is applied to BE, and its performance is evaluated and compared with the existing methods. Furthermore, we incorporate our recently developed strategy for model selection with incomplete data, known as the E-MS algorithm, with the RF to address the common missing value concerns in BE. Our study reveals some interesting findings in association with the missing data mechanisms. The proposed method is illustrated with a real data analysis involving QTL mapping for an agricultural study on barley grains.
This work is joint with Jiming Jiang of the University of California, Davis, and J. Sunil Rao of the University of Miami.

High dimensional cluster feature selection

Zhen PANG
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
E-mail: zhen.pang@polyu.edu.hk

Keywords: Variable selection; Variable screening; SIS; Elastic net

Abstract: This talk concerns with variable screening when highly correlated variables exist in high dimensional linear models. We propose a novel cluster feature selection procedure based on the elastic net and linear correlation variable screening to enjoy the benefits of the two methods. When calculating the correlation between the predictor and the response, we consider the highly correlated group of the predictors instead of the individual ones. This is in contrast to the usual linear correlation variable screening. Within each correlated group, we apply the elastic net to select and estimate the variables. This avoids the drawback of mistakenly eliminating true non-zero coefficients for highly correlated variables like LASSO (Tibshirani, 1996) does. After our

procedure, maximum absolute sample correlation coefficient between clusters becomes smaller and any common model selection methods like SIS (Fan and Lv, 2008) or LASSO can be applied to improve the results. Extensive numerical examples including pure simulation examples and semi-real examples are conducted to show the good performances of our procedure.

IS34 MODELING AND ANALYSIS OF COMPLEX BIOMEDICAL DATA

Session Organizer and Chair: Donguk Kim, Sungkyunkwan University, South Korea
Venue: Stephen Riady Global Learning Room
Time: 17 Dec, 13:40-15:40

Penalized exponential tilt model for analysis of high-dimensional DNA methylation data

Hokeun Sun
Department of Statistics, Pusan National University, South Korea
E-mail: hsun@pusan.ac.kr

Keywords: Exponential tilt model, DNA methylation, High-dimensional data, Network-based regularization

Abstract: In epigenetic studies of human diseases, it has been common to compare DNA methylation levels between cancer tissues and normal tissues to identify cancer-related genetic sites. For case-control association studies with high-dimensional DNA methylation data, a network-based penalized logistic regression has been proposed in our earlier article. Network regularization is very efficient for analysis of highly correlated methylation data. However, recent studies found that the methylation levels of the cancer and normal tissues could differ not only in means but also in variances. Penalized logistic regression has a limitation to detect any differences in variances. In this article, we introduce a penalized exponential tilt model using network-based regularization and demonstrate that it can identify differentially methylated loci between cancer and normal tissues when their methylation levels are different in means only, variances only or in both means and variances. We also applied the proposed method to a real methylation data from an ovarian cancer study where methylation levels over 20,000 CpG sites were generated from Illumina Infinium HumanMethylation27K Beadchip. We identified additional methylation loci that were missed by the penalized logistic regression.

Effect-size distributions of human complex traits and diseases from genome-wide association studies

Ju-Hyun Park
Department of Statistics, Dongguk University-Seoul, South Korea
E-mail: juhyunp@dongguk.edu

Keywords: genome-wide association study; effect size; distribution estimation; risk prediction

Abstract: Although recent genome-wide association studies have led to the discoveries of many susceptibility loci, much of the heritability of the individual traits and diseases remain unexplained. Many different approaches to explaining 'missing heritability' have been proposed with different types of variants. As a recent tool for genome-wide complex trait analysis showed that a moderate amount of the heritability could be explained by common variants, we assess the effect-size distribution of common susceptibility loci by accounting for power for detecting known loci from their original genome-wide studies. With the estimated effect-size distributions for various traits and diseases such as adult height, high-density lipoprotein(HDL), Type 2 diabetes, and Crohn's diseases, we will illustrate how useful and important it is to know the effect-size distribution for various human complex traits and diseases for planning future genome-wide association studies and for projecting risk prediction.

Fitting semiparametric accelerated failure time models for nested case-control data

Sangwook Kang
Department of Applied Statistics, Yonsei University, Seoul, Korea
E-mail: kanggi1@yonsei.ac.kr

Keywords: Gehan weight, Induced smoothing, Survival analysis, Weighted log-rank test

Abstract: A nested case-control study is an efficient cohort-sampling design in which a subset of controls are sampled from the risk set at each event time. Since covariate measurements are taken only for the sampled subjects, time and efforts of conducting a full scale cohort study can be saved. In this paper, we consider fitting a semiparametric accelerated failure time (AFT) model to failure time data from a nested case-control study. We propose to employ an efficient induced smoothing procedure for rank-based estimating method for regression parameter estimation. For variance estimation, we propose to use an efficient resampling method that is based on the sandwich form of the asymptotic variance. We extend our proposed methods to a generalized nested case-control study that allows a sampling of cases. Finite sample properties of the proposed estimators are investigated via an extensive stimulation study. An application to a tumor study illustrates the utility of the proposed method in routine data analysis.

Multiple imputation for competing risks survival data via pseudo-observations

Seungbong Han
Department of Applied Statistics, Gachon University
E-mail:hanseungbong@gmail.com

Keywords: Competing Risks, Hepatocellular Carcinoma, Missing Data, Multiple Imputation, Random Forest.

Abstract: In biomedical research, competing risks are commonly encountered. Regression models for competing risks data can be developed based on data routinely collected in hospitals or general practices. However, these data sets usually contain missing values for the covariates. To overcome this problem, multiple imputation is often used to fit regression models under a missing-at-random assumption. Here, we introduce a multivariate imputation in chained equations (MICE) algorithm to deal with competing risks survival data. Using pseudo-observations, we make better use of the available outcome information by accommodating the competing risk structure. Lastly, we illustrate the practical advantages of our approach using simulations and an example based on hepatocellular carcinoma data.

IS35 ADVANCED PARAMETRIC AND NONPARAMETIC STATISTICAL APPROACHES

Session Organizer and Chair: Yung-Seop Lee, Dongguk University, South Korea
Venue: Seminar Room 1
Time: 17 Dec, 13:40-15:40

A class of rectangle-screened multivariate normal distributions and its applications

Hyoung-Moon Kimb
Department of Applied Statistics, Konkuk University, Seoul, Korea

Keywords: Truncated multivariate normal distribution; Rectangle-screened multivariate distribution; Closure Property; Applications.

Abstract: A screening problem is tackled by proposing a parametric class of distributions designed to match the behavior of the partially observed screened data. This class is obtained from the nontruncated marginal of the rectangle-truncated multivariate normal distributions. Motivations for the screened distribution as well as some of the basic properties, such as its characteristic function, are presented. These allow us a detailed exploration of other important properties that include closure property in linear transformation, in marginal and conditional operations, and in a mixture operation as well as the first two moments and some sampling distributions.

Various applications of these results to the statistical modeling and data analysis are also provided.

Automated K-means clustering

Sung-Soo Kim
Department of Information Statistics, Korea National Open University
E-mail: sskim@knou.ac.kr

Keywords: Automated K-means clustering, variable selection, outliers, VS-KM, adjusted rand index, Mahalanobis distance.

Abstract: Two crucial problems of K-means clustering are deciding the number of clusters and initial centroids of clusters. Also variable selection and outlier identification are important tasks. We provide automated k-means clustering process and provide R implentation program. The Automated K-means clustering procedure consists of three processes: (i) automatically calculating the cluster number and initial cluster center whenever a new variable is added, (ii) identifying outliers for each cluster depending on used variables, (iii) selecting variables defining cluster structure in a forward manner. To decide the number of clusters, we used Mojena's rule combined with the Ward's methods using the sampled data. To select variables, we applied VS-KM (variable-selection heuristic for K-means clustering) procedure (Brusco and Cradit, 2001). To identify outliers, we used a hybrid approach combining a clustering based approach and distance based approach. Simulation results indicate that the proposed automated K-means clustering procedure can be useful for large data sets and effective to select variables and identify outliers.

Sparse HDLSS discrimination with constrained data piling

Yongho Jeon
Yonsei University, South Korea

Abstract: Regularization is a key component in high dimensional data analyses. In high dimensional discrimination with binary classes, the phenomenon of data piling occurs when the projection of data onto a discriminant vector is dichotomous, one for each class. Regularizing the degree of data piling yields a new class of discrimination rules for high dimension low sample size data. A discrimination method that regularizes the degree of data piling while achieving sparsity is pro- posed and solved via a linear programming. Computational efficiency is further improved by a sign-preserving regularization that forces the signs of the estimator to be the same as the mean difference. The proposed classifier shows competitive performances for simulated and real data examples including speech recognition and gene expressions.

Bayeisan nonparametric models with species sampling priors

Jaeyong Lee
Department of Statistics, Seoul National University, South Korea

Abstract: The species sampling model is a discrete random probability distribution represented as the sum of the random support points with random weights. We consider theoretical background of the species sampling models and inference of spatially varying densities based on mixtures of dependent species sampling models. The spatial dependency is introduced by modeling the weights through the conditional autoregressive model. The proposed models are illustrated in two simulated data sets and show better performance than the density estimation methods for which the dependency is not incorporated. The proposed method is also applied to Climate Prediction Center Merged Analysis of Precipitation (CMAP) data of 33 years over Korea. The probability density functions of the precipitation over grid points are estimated.

IS36 DATA REPRESENTATION FOR ANALYZING BIG DATA

Session Organizer and Chair: Junji Nakano, The Institute of Statistical Mathematics, Japan
Venue: Stephen Riady Global Learning Room
Time: 18 Dec, 10:20-12:20

Multidimensional scaling for one-mode three-way symbolic dissimilarity data

Kimihiro Hasegawa
Graduate School of Culture and Information Science, Doshisha University, Japan
E-mail: dio0004@mail4.doshisha.ac.jp

Keywords: Triadic distance; Interval-valued dissimilarity data; Histogram-valued dissimilarity data

Abstract: One-mode three-way dissimilarity data are defined as dissimilarity data among three objects. These data are provided by respondents, which simultaneously evaluate the relationships among the three items. In this study, it is important to visually interpret these relationships. De Rooji and Gower (2003) proposed the use of multidimensional scaling(MDS) on the basis of a triadic-distance model. This model's advantage is that it easily interprets geometrical representations. On the other hand, concepts of individuals has attracted much attention by researchers (Bock and Diday, 2000). In our analysis, if the dissimilarity data are described as single values by aggregating the data referring to the three objects, information on their variability is lost. Therefore, interval-valued dissimilarity data are required to provide additional descriptive parameters for the symbolic dissimilarity data. Here, the application of a triadic-distance model is extended from single-

valued to symbolic data, and a new MDS is proposed for one-mode three-way dissimilarity data on the basis of the model of Denoeux and Masson (2000).

Visualizing dissimilarity among aggregated symbolic data

Nobuo Shimizu
The Institute of Statistical Mathematics, Japan
Email: nobuo@ism.ac.jp

Keywords: Aggregated symbolic data, Clustering, Likelihood ratio test statistics

Abstract: In these days, huge amount of individual data are frequently available in all fields of science, engineering and social activities. Such \big data" require new data representation and analysis methods for being understood by analysts. Symbolic data analysis provides techniques for handling them. Traditional symbolic data analysis uses information only about marginal distribution of each variable. We consider that individual data can be divided into some naturally defined groups and be expressed by up to second order moments which include information about both marginal distributions and a joint distribution of each group. We also consider that individual data consists of continuous and categorical variables. Then we use means, variances and covariances for summarizing continuous variables, and contingency tables for summarizing categorical variables. We call them \aggregated symbolic data (ASD)".
For clustering such ASD, we define a dissimilarity measure based on likelihood ratio test statistics and their decomposition to interpret it easily. We also propose visualizing them for grasping their characteristics intuitively. Example data are analyzed by the proposed method.

Empirical study on analytic software toward 'Mini-data' analysis

Hiroyuki Minami and Masahiro Mizuta
Information Initiative Center, Hokkaido University, Japan
E-mail: mizuta@iic.hokudai.ac.jp

Keywords: Computational Statistics, Big data, Cloud computing

Abstract: We offer some reviews on some new analytic environment and libraries and discuss the extensibility to thoughtful data analysis in comparison with the language R.
The word 'Big data' is in the public now and we, statisticians, should face them in real data analysis. However, most are miscellaneous and provided in ugly format. If we would do the appropriate analysis, adequate shaping and shrinking should be recommended. We propose the adequate intermediate data 'Mini-data'. To get them, some statistical methods to the original ugly data (e.g. random sampling, dimension reduction, collective approaches including Functional Data Analysis and Symbolic Data Analysis) occur to us,

but simple computer-oriented methods like an elimination for the duplicated data are also required, mainly from the operational view. The language R is popular and has powerful potential in statistics, but due to the limitation derived from its original design, it might not be suitable for a large amount of data in real analysis. Now, many software libraries available in distributed computer environment with some 'glue' languages including Python are introduced. However, we are afraid that not all statisticians are familiar with them and there are few empirical studies to discuss the topics. In the study, we first give the reviews of the libraries in Python (including Numpy, SciPy and matplotlib) and discuss the comparison with R, through a practical example using the big data of radio dose ratios in Japan.

A study on environmental radioactivity level data with functional data analysis

Satoshi Kikuchi
Graduate School of Information Science and Technology, Hokkaido University, Japan
E-mail: kikuchi@iic.hokudai.ac.jp

Keywords: Air dose rate, monitoring post

Abstract: 1. we analyze environmental radioactivity level data with functional data analysis (FDA) to predict transition of air dose rate.
2. The Fukushima Daiichi nuclear accident occurred in March 2011. Nuclear Regulation Authority measures air dose rate in many places and various methods. Air dose rates have been generally reduced, and we would like to know the transitions. It is hard to predict them since they do not follow a simple physical model. It is important to turn out the specific features in environmental radioactivity level data for much accurate prediction.
3. We analyze monitoring post data. Monitoring posts measure Gamma ray every 10 minutes at about 3000 sites in Fukushima prefecture. We can regard them as functional data and apply FDA methods to the data. As a result, we detected some remarkable features including the seasonal variation. Air dose rates have been decreased more than gamma attenuation in most sites, and the reduction rates are not constant. They tend to decrease in summer and increase in winter.

IS37 INTERPRETING THE CONSUMER: FROM COMMUNICATION TO CUSTOMIZATION
Session Organizer: Tim Banks, The Nielsen Company, Singapore
Session Chair: Whye Loon Tung, The Nielsen Company, Singapore
Venue: Seminar Room 1
Time: 18 Dec, 10:20-12:20

Automatically building stories from news: from the morosini codex to online news feeds

Siew Ann Cheong and Andrea Nanetti
School of Physics and Mathematical Sciences, Nanyang Technological University, Singapore
Email: cheongsa@ntu.edu.sg

Keywords: Data Mining; Automatic Narratives

Abstract: Besides the rise of transactional big data, we also face a torrent of news feeds from formal and informal sources online. Making sense of this information stream requires more than traditional data mining to find patterns or correlations, because the human mind is simply not good at digesting such low-level meta-information. Instead, the human mind feeds on stories, which are coherent collections of narratives comprising such elements as who, what, when, where, why, how. In the past, we humans struggle to discover these coherent storylines from a sparse set of news that offers very little coverage, and even then it is possible to find multiple storylines that are internally quasi-coherent, but conflicts with each other. In the present, with hundreds and thousands of news arriving every day, this problem of overlapping storylines is far worse than most people can imagine. In this talk, we will describe our work extracting key narrative elements automatically from a historical data set (the Morosini Codex). We will describe how a human expert can prepare a curated set of important actors in the historical data, along with a curated set of important interactions. This allows us to construct a time-integrated actor-to-actor complex network. Following this, we examine the patterns the links are activated in sliding time windows, to identify the key periods and key events, and therefrom the key actors, key locations, and key motives. We illustrate this procedure for generating automatic narratives using a few examples, before moving on to suggest that the same method can be applied to online news feeds. The output from this novel form of data mining is in the form of a digest, which recounts only the most salient storyline for a given list of actors and actions, i.e. for an investment trader, the actors are stocks, while the actions are different market events, and the digest is a summary outlook on investor sentiment, whereas for modern national governments, the actors are other governments and rogue organizations, while the actions are their postures on each other, and the digest is a summary on the security outlook for each actor.

Profiling latent user attributes in social media

Richard Oentaryo
Living Analytics Research Centre, School of Information Systems, Singapore Management University, Singapore
Email: roentaryo@smu.edu.sg

Keywords: Latent Attributes, Social Media, Profiling

Abstract: In recent years, we have witnessed a dramatic growth in human activities taking place in social media such as Twitter and Facebook. This has led to a massive generation of digital data about user behaviors. The availability of such data has sparked the desire to learn more about users as well as organizations, fuelling in turn the emergence of new services for peer interaction, marketing, and content sharing. The ability to profile user preferences and attributes thus has important applications in personalization, advertising, and recommendation. Despite the abundance of user-generated data, meta-information about personal attributes that are directly useful for personalized services and recommendations are often not available. In Twitter, for instance, users rarely provide such demographic information as gender, age, religion, or marital status. Nevertheless, recent studies have shown that it is possible to use statistical learning methods to infer these attributes, based on data traces (e.g., users' contents and social ties) that users revealed in social media. In this work, we present empirical studies on profiling the latent user attributes in Twitter by leveraging statistical learning approaches. We present a generic profiling framework for inferring user attributes, based upon which comprehensive and systematic investigation on a rich variety of features and the general user population can be carried out. In addition to generic feature transformation pipelines, the framework features a classifier bank that consists of statistical learning methods that are accurate/robust and facilitate good data understanding such as the importance of different features. We describe the application of our framework to two novel user profiling tasks, namely account type and bot behavior classification.

Business location analytics using social media

Jovian Lin and Richard Oentaryo
Living Analytics Research Centre, School of Information Systems, Singapore Management University, Singapore
Email: jovianlin@smu.edu.sg, roentaryo@smu.edu.sg

Keywords: Gradient Boosting, Recommender Systems

Abstract: Location is a vital aspect of retail success, especially for brick-and-mortar stores, as 94% of retail sales are still in physical stores. To optimize the likelihood of success for their stores, business owners require the knowledge of not only where their potential customers are, but also their surrounding competitors and potential allies, which may be stores that offer complementary services or help to draw in customer crowds. However, assessing and picking a store location is a herculean task for the business owner. To carry out the above tasks well, there are many factors to be considered and each factor requires gathering and analyzing the relevant data. In this work, we investigate the use of Facebook's check-ins to evaluate or estimate the success of businesses. Using a dataset of more than 20,000 food businesses on Facebook pages, we conduct analysis of several success-related factors including business categories, locations and neighboring businesses. From these factors, we extract a set of relevant features and develop a prediction model that estimates a business's success.

This prediction model uses a robust statistical learning method called gradient boosting (GB) to combine six different types engineered features. We compare this method with other baselines including distance-based nearest neighbors and support vector regression (SVR), and show how GB significantly outperforms all the other methods. On top of that, our experiments have shown that the success of neighboring business contributes the key features to perform accurate prediction. We finally illustrate the application of such a prediction method using a user-friendly food business recommender system.

Correlation analysis of social signals and nonverbal audio-video features

Justin Dauwels
School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore
Email: jdauwels@ntu.edu.sg

Keywords: Social behaviour, realtime feedback systems

Abstract: Understanding of human social behaviour is key to building socially intelligent computer systems for the modern world. Such social behaviour is often manifested through nonverbal cues ingrained in social indicators such as Dominance, Politeness and Confusion used during interactions.
The field offers excellent opportunities for the application of advanced statistical learning techniques given the massive data volumes collected by instrumentation. Rapid processing of these data streams also presents interesting challenges for real-time feedback systems. This paper reviews recent work at NTU.
Annotations of the social indicators from different judges on a likert scale are used for training machine learning classifiers. A linear correlation study is presented to investigate the relation between ten of such social indicators and between the indicators and the nonverbal audio-video features extracted from a total of 496 two-person face-to-face recordings. Results indicate redundancy of information, with many of the indicators strongly correlated to others. However, indicators such as Confusion are not related to any other indicator, suggesting it to be an independent indicator. In addition, each social indicator is only correlated to a subset of the nonverbal audio-video feature cues. Such redundancies can lead to a simpler implementation of a real-time socio-feedback system.

Application of analytics to massive data and business solutions

Ashok Charan
Marketing Analytics, Business School, National University of Singapore, Singapore
Email: bizakc@nus.edu.sg

Keywords: Business Analytics, Consulting

Abstract: Consumer Analytics: Analytic methods for analysis of continuous household purchases data. The data would usually pertain to consumer transactions (e.g. online transactions, outlet transactions), loyalty panels, consumer panels, credit card usage, telecommunication services usage and a host of other application areas.
Retail Analytics: Analytic methods for analysis of continuous outlet level sales data. The data is usually sourced from point-of-sales scan terminals at stores, or from retail audits conducted by research firms.
The discussion on both these topics would pertain to methods and techniques used by practitioners to understand their consumers, and identify and resolve marketing issues. The presentation will cover how analytics and data fusion are applied to business problems, giving client companies the `analytic edge' over their competition. Prof. Ashok has both industrial and academic experience, and will include practical advice on statistical consulting to Industry.

IS38 RISK ANALYTICS

Session Organizer and Chair: Stefan Lessmann, School of Business and Economics, Humboldt-University of Berlin, Germany
Venue: Seminar Room 1
Time: 17 Dec, 16:00-18:00

Estimating the profitability of transactors and revolvers

Hsin-Vonn Seow
Nottingham University Business School, University of Nottingham-Malaysia Campus, Malaysia
E-mail: Hsin-Vonn.Seow@nottingham.edu.my

Keywords: Credit Scoring; Profitability Scoring; Survival Analysis.

Abstract: In consumer lending, credit decisions are becoming increasingly more complex. The conventional approach is to develop a credit risk-based model to help with the decision of how to classify customers as good (those with low risk of defaulting) or bad (those with high risk of defaulting). This approach enables lenders to make a decision on whom to offer credit to. Credit card customers can be further classified into transactors and revolvers. Transactors are defined as customers that pay off their balance every month whilst revolvers are those who do not consistently pay off their monthly balance but tend to carry some of it over to the next time period. Profit generated by transactors and revolvers are quite different. For revolvers, the profit mainly comes from the interest charged on the carrying balance. For transactors, it comes from the merchandise fee. In addition, the spending and attrition behaviour of transactors/revolvers are quite different. Therefore, it seems logical to estimate the profit of these two groups of credit cardholders

separately. Using a large credit card dataset provided by a lender, we are going to conduct computational analyses to estimate the profitability of transactors and revolvers. These will involve the use of survival analysis to look at when a transactor first becomes a revolver, when a revolver is going to default and when a transactor/revolver is going to churn. In addition, a regression model will be used to estimate the average purchases of transactors/revolvers. Using these results, a model will then be developed to calculate the profitability of different types of cardholders more accurately. The proposed approach could be used by lenders to develop a profit scoring model for making credit decisions.

Credit risk modeling: a benchmark of machine learning methods

Stefan Lessmann
School of Business and Economics, Humboldt-University of Berlin.
E-mail: stefan.lessmann@hu-berlin.de

Keywords: Credit Scoring; Predictive Analytics; Benchmark.

Abstract: Credit scoring involves the development of empirical models to support decision making in the retail credit business. In particular, financial institutions use predictive models, called scorecards, to decide upon new credit applications or to manage existing loan agreements. Such score-cards estimate the probability of a borrower to default. The paper presents results of a comprehensive benchmark experiment in which we contrast the performance of 41 alternative forecasting methods across eight real-world credit scoring data sets. The objective of the study is to examine the potential of recent advancements in the field of predictive analytics for credit scoring. More specifically, we consider several novel learning methods that are for the first time assessed in a credit scoring context. Furthermore, using the principles of cost-sensitive learning, we shed light on the link between the (statistical) accuracy of scorecard predictions and a scorecard's business value. Finally, we examine the extent to which different indicators of predictive accuracy agree in their scorecard assessment and thereby contribute to a recent debate concerning the appropriateness of receiver operating characteristics. Our study provides valuable insight for professionals and academics in credit scoring. It helps practitioners to stay abreast of technical advancements in predictive modeling. From an academic point of view, the study provides an independent assessment of recent prediction methods and offers a new baseline to which future models can be compared.

Recovery of foreign interest rates from exchange options

Yasushi Ota
Doshisha University, Japan

Abstract: In currency markets practitioners take a strategy in which they sell a certain currency with a relatively low interest rate and buy a different currency yielding a higher interest rate and attempt to capture the difference between the rates.

This strategy is often called the carry tread. So, it motivates the carry trader to know valuable information between interest and exchange rates for currencies. The stochastic behavior St at time t of a risky asset such as exchange rate is modelled by

$$dS_t = \mu(t, S_t)S_t dt + \sigma(t, S_t)S_t dW_t, \; S_0 = x \quad (1.1)$$

where the process Wt is the Brownian motion. The parameters μ(t, x) and σ(t, x) are called the real drift and the local volatility of the underlying asset, respectively.

In the finance theory (see [1]) on the no–arbitrage market, it is well–known that the behavior of financial derivative u(t, x) satisfies the following PDE:

where r is the instantaneous return of the riskless asset such as domestic interest rate for currency. Here the parameter m(t, x), which is not always equal to r(t, x), is implied by the no–arbitrage condition.

Dupire[4] first found that market price of call options given for all possible strikes K and maturities T completely determine the local volatility by using the dual equation to the above PDE with $m(t, x) \equiv m$, where m is constant.

Unfortunately, options data are typically available only for a single maturity, in particular, in the foreign–exchange markets. In Bouchouev and Isakov[2], Isakov and Valdivia[3], in the case $\sigma(T, K) \equiv \sigma(K)$ they formulated the inverse problem identifying $\sigma(K)$ such that the last PDE follows with

$$u(T, K)|_{T=t} = \max\{0, x - K\}, \quad (1.4)$$
$$u(T, K)|_{T=T^*} = u^*(K),$$

where $u^*(K)$ is the market price of call options with different strikes K for a given maturity T^*.

In this talk, we consider the foreign – exchange rate model in which the volatility of the behaviour (1.1) of the exchange rate is positive constant, i.e. $\sigma(t, x) \equiv \sigma$ and its drift is time – independent. Then, a price u(t, x) for option with strike K and maturity T satisfies (1.3) and (1.4) where m(t, x) is time–independent, i.e. m(t, x) = m(x) and is equal to the difference between domestic and foreign interest rates. Under the assumption that we know σ and r, from market prices of binary options with different strikes K and a single maturity T^* we attempt to identify m(K).

In this talk, we first explain our problem and formulate the inverse problem of the option pricing. The difficulty with the original problem suggests studying a standard linearization procedure, which is the approximate algorithm neglecting terms of higher order with respect to a perturbation. Second, we use the linearization and derive the integral equation for solving approximately this problem. Third, we discretize this integral equation and provides several numerical examples of recovering foreign interest rates. For a given drift m(K) we generate data by solving PDE and by using the generated data we recover the drift m(K) from the integral equation. Finally, by using our integral equation, we try the recovering foreign interest rates from a real market data.

IS39 PRACTICAL APPROACHES TO PROBLEMS IN COMPUTATIONAL STATISTICS

Session Organizer and Chair: John Ormerod, University of Sydney, Australia
Venue: LT51
Time: 17 Dec, 13:40-15:40

Classified mixed model prediction

Jiming Jiang
University of California, Davis, USA

Abstract: Many practical problems are related to prediction, where the main interest is at subject (e.g., online shopper) or (small) sub-population (e.g., small community) level. In such cases, it is possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to. This way, the new subject is potentially associated with a random effect corresponding to the same class in the training data, so that method of mixed model prediction can be used to make the best prediction. We propose a new method, called classified mixed model prediction (CMMP), to achieve this goal. We develop CMMP for both prediction of mixed effects and prediction of future observations, and consider different scenarios where there may or may not be a ``match'' of the new subject among the training-data subjects. Theoretical and empirical studies are carried out to study the properties of CMMP, and its comparison with existing methods. In particular, we show that, even if the actual match does not exist between the group of the new observations and those of the training data, CMMP still helps in improving prediction accuracy. A real-data application is considered. This work is joint with J. Sunil Rao, Jie Chen of the University of Miami, USA, and Thuan Nguyen of the Oregon Health \& Science University, USA.

Variational inference for Bayesian semiparametric additive models

Heng Lian
University of New South Wales, Australia

Abstract: We develop a mean field variational Bayes approximation algorithm for posterior inferences of partially linear additive models with simultaneous and automatic variable selection and linear/nonlinear component identification abilities. To solve the problem induced by some complicated expectation evaluations, we tried two approximations based on Monte Carlo method and Laplace approximation respectively. With high accuracy, the algorithm we derived is much more computationally efficient than the existing Markov Chain Monte Carlo (MCMC) method. The simulation examples are used to demonstrate the performance of our new algorithm versus MCMC. The proposed approach is further illustrated on a real dataset.

A collapsed variational approach to Bayesian model selection

John Ormerod
University of Sydney, Australia

Abstract: In recent years a great deal of research has been conducted on model selection. Most of this has focused on (1) penalized regression methods or (2) Markov Chain Monte Carlo (MCMC). The variational Bayes approach proposed by You et al (2015) stands as an alternative method which is fast like penalized regression approaches with competitive model selection accuracy when compared to MCMC. In this talk we propose an improvement upon this work which seeks to relax the posterior independence assumptions of You et al (2015) resulting in improved inferences and an interpretable structure from which various relationships between variables can be inferred.

Sequential Monte Carlo methods for Bayesian elliptic inverse problems

Muzaffer Ege Alper
National University of Singapore, Singapore

In this talk we consider a Bayesian inverse problem associated to elliptic partial differential equations (PDEs) in two and three dimensions. This class of inverse problems is important in applications such as hydrology, but the complexity of the link function between unknown field and measurements can make it difficult to draw inference from the associated posterior. We prove that for this inverse problem a basic SMC method has a Monte Carlo rate of convergence with constants which are independent of the dimension of the discretization of the problem; indeed convergence of the SMC method is established in a function space setting. We also develop an enhancement of the sequential Monte Carlo (SMC) methods for inverse problems which were introduced in Kantas, et al (2014); the enhancement is designed to deal with the additional complexity of this elliptic inverse problem. The efficacy of the methodology, and its desirable theoretical properties, are demonstrated on numerical examples in both two and three dimensions.

IS40 RECENT DEVELOPMENTS IN APPLIED ECONOMETRICS AND FINANCE
Session Organizer and Chair: Phuong Anh Nguyen, International University - VNUHCM, Vietnam
Venue: Seminar Room 2
Time: 17 Dec, 10:40-12:40

Nonparametric versus parametric choice models: an empirical investigation

Michel Simioni
INRA, UMR MOISA, Montpellier, France

Abstract: Recent developments in the nonparametric estimation of conditional probability distribution functions (PDFs) offers practitioners a flexible framework for estimation and inference.
The modelling of conditional PDFs can be extremely useful for a range of tasks including direct quantile estimation and prediction of consumer choice, by way of examples. In this paper we assess the potential of this nonparametric estimator for improved modelling of consumer choice. We model a dataset in which the outcome is a binary consumer choice while the covariates consist of both continuous and categorical (discrete) variables. We assess the relative performance of the nonparametric estimator and the parametric Probit specification that dominates in applied settings. We compare these estimators using a variety of measures and also assess their performance on independent data drawn from the same underlying distribution and test for significant differences. Finally, we demonstrate that the nonparametric estimator reveals certain features present in the data that lie undetected by the parametric Probit model.

Factorisable sparse tail event curves

Wolfgang K. Härdle
Humboldt University of Berlin, Germany

Abstract. In this paper, we propose a multivariate quantile regression method which enables localized analysis on conditional quantiles and global comovement analysis on conditional ranges for high-dimensional data. The proposed method, hereafter referred to as FActorisable Sparse Tail Event Curves, or FASTEC for short, exploits the potential factor structure of multivariate conditional quantiles through nuclear norm regularization and is particularly suitable for dealing with extreme quantiles. We study both theoretical properties and computational aspects of the estimating procedure for FASTEC. In particular, we derive nonasymptotic oracle bounds for the estimation error, and develop an efficient proximal gradient algorithm for the non-smooth optimization problem incurred in our estimating procedure. Merits of the proposed methodology are further demonstrated through applications to

Conditional Autoregressive Value-at-Risk (CAViaR) (Engle and Manganelli; 2004), and a Chinese temperature dataset.

Financial ratios and efficiency of Vietnamese commercial banks: a semiparametric approach

Phuong Anh Nguyen
International University - VNUHCM, Vietnam

Abstract. The assessment of bank performance is a much debated issue in the literature on banking. This paper contributes to the debate by comparing performance assessment through financial ratios and Data Envelopment Analysis. Recent bootstrapping techniques are used to see what financial ratios can explain about bank efficiency, depending on the chosen approach of bank technology: value-added or intermediation.

IS41 COMPUTING, SELECTION AND REDUCTION IN HIGH-DIMENSIONAL STATISTICAL METHODS

Session Organizer and Chair: Xin Zhang, Department of Statistics, Florida State University, USA.
Venue: Seminar Room 2
Time: 17 Dec, 13:40-15:40

Conformational sampling and energy evaluation of protein loops

Samuel W.K. Wong
Department of Statistics, University of Florida, USA
E-mail: swkwong@stat.ufl.edu

Keywords: protein structure prediction, sequential Monte Carlo, loop modeling, conformational sampling, energy functions.

Abstract: Using computation to predict a protein's 3D structure from its amino acid sequence remains a highly challenging problem in biology. This talk focuses on one particular subproblem: the structure prediction of the variable loop regions in proteins that connect the more ordered helices and sheets. This requires two main ingredients: a sampling algorithm, and an energy function. We will overview a recently developed method, inspired by sequential Monte Carlo techniques, to tackle the task of efficiently sampling the space of low-energy conformations in the loop region. An ideal energy function should then assign the lowest energy to the conformation that is closest to the truth, but in practice this is difficult to achieve. We use our method to systematically evaluate some commonly used energy functions and gain insights about the energy landscape of the loop conformational space.

A unified algorithm for fitting penalized models with high dimensional data

Yi Yang
McGill University, Canada

Abstract: In the light of high dimensional problems, research on the penalized model has received much interest. Correspondingly, several algorithms have been developed for solving penalized high dimensional models. In this thesis, we propose fast and efficient unified algorithms for computing the solution path for a collection of penalized models. In particular, we study the algorithm for solving ℓ1 penalized learning problems and the algorithm for solving group-lasso learning problems. These algorithm take advantage of a majorization-minimization trick to make each update simple and efficient. The algorithms also enjoy a proven convergence property. To demonstrate the generality of our algorithms, we further extend these algorithms on a class of elastic net penalized large margin classification methods and the elastic net penalized Cox's proportional hazards model. These algorithms have been implemented in three R packages gglasso, gcdnet and fastcox, which are publicly available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages. On simulated and real data, our algorithms consistently outperform the existing software in speed for computing penalized models and often delivers better quality solutions.

Envelopes and parsimonious tensor response regression

Xin Zhang
Department of Statistics, Florida State University, USA
E-mail: henry@stat.fsu.edu

Keywords: Envelope method; multidimensional array; multivariate linear regression; reduced rank regression; sparsity principle; tensor regression.

Abstract: Aiming at abundant scientific and engineering data with not only high dimensionality but also complex structure, we study the regression problem with a multi-dimensional array (tensor) response and a vector predictor. Applications include, among others, comparing tensor images across groups after adjusting for additional covariates, which is of central interest in neuroimaging analysis. We propose parsimonious tensor response regression adopting a generalized sparsity principle. It models all voxels of the tensor response jointly, while accounting for the inherent structural information among the voxels. It effectively reduces the number of free parameters, leading to feasible computation and improved interpretation. We achieve model estimation through a nascent technique called the envelope method, which identifies the immaterial information and focuses the estimation based upon the material information in the tensor response. We demonstrate that the resulting estimator is asymptotically efficient, and it enjoys a competitive finite sample performance. We also illustrate the new method on two real neuroimaging studies.

(Joint work with Dr. Lexin Li)

On marginal sliced inverse regression for ultrahigh dimensional model free feature selection

Zhou Yu
School of Finance and Statistics, East China Normal University, Shanghai, China
E-mail: zyu@stat.ecnu.edu.cn

Keywords: marginal coordinate test; sliced inverse regression; sufficient dimension reduction; sure independence screening.

Abstract: Model-free variable selection has been implemented under the sufficient dimension reduction framework since the seminal paper of Cook (2004). In this paper, we extend the marginal coordinate test for sliced inverse regression (SIR) in Cook (2004) and propose a novel marginal SIR utility for the purpose of ultrahigh dimensional feature selection. Two distinct procedures, Dantzig selector and sparse precision matrix estimation, are incorporated to get two versions of sample level marginal SIR utilities. Both procedures lead to model-free variable selection consistency with predictor dimensionality p diverging at an exponential rate of the sample size n. As a special case of marginal SIR, we ignore the correlation among the predictors and propose marginal independence SIR. Marginal independence SIR is closely related to many existing independence screening procedures in the literature, and achieves model-free screening consistency in the ultra-high dimensional setting. The finite sample performances of the proposed procedures are studied through synthetic examples and an application to the small round blue cell tumors data.

IS42 PRECISION MEDICINE: FROM BENCHSIDE TO BEDSIDE
Session Organizer and Chair: Ying Yuan, University of Texas MD Anderson Cancer Center, USA
Venue: Stephen Riady Global Learning Room
Time: 19 Dec, 9:00-11:00

Incorporating historical data from a previous study to design a dose-finding trial in a new region

Satoshi Morita
Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan

Abstract: In oncology area, following a phase I dose-finding trial completed in a certain population of patients, further phase I trials are often conducted to determine the maximum tolerated dose (MTD) for different patient

subpopulations. This may be due to concerns about possible differences in treatment tolerability between subgroups. We propose a Bayesian clinical trial design incorporating historical data to design a dose-finding trial in a new patient population. Our proposed approach aims to appropriately borrow strength from a previous trial to improve the MTD determination in a new trial. We propose a historical-to-current parameter to evaluate the similarity in dose-toxicity relationship between two patient subgroups, e.g., Caucasian and Asian patient populations. We present a simulation study of our proposed method to explore its properties. The results of a simulation study to examine the performance of our proposed method are encouraging.

Somatic DNA copy number aberrations in the peripheral blood of patients with solid tumors

Li Zhang
Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, USA

ABSTRACT: Although DNA somatic copy number alterations (SCNAs) are known to play a pivotal role in cancer, SCNAs in the blood of cancer patients have not been systemically characterized. Here, we performed a genome-wide analysis of focal SCNAs in the peripheral blood of 8,870 cancer patients, as well as in the histologically “normal” tissue and tumor specimens. 69 genomic sites were discovered to contain significantly higher frequency of SCNAs in blood than that in tumor and normal tissue. In individual patients, SCNAs in some of these sites occurred exclusively in blood but absent in tumor and normal tissue. 13% patients harbored at least one SCNA of the most frequent sites in blood. The SCNAs enriched in blood were correlated with increased age, shorter progression free survival of the patients and increased immune activity in the primary tumors. These results suggest the SCNAs in blood represent systemic, tumor-extrinsic genomic aberrations that alter the immune system affecting local immune responses in the tumor. The genes encoded in the SCNAs will presumably generate new insights and foster the further development of precision diagnostics and targeted therapies.

MOST: Multi-stage optimal sequential trial design for phase ii clinical trials with biomarker subgroups

Ying Yuan
Department of Biostatistics & University of Texas MD Anderson Cancer Center, USA
E-mail: yyuan@mdanderson.org

Keywords: subgroups; personalized medicine; precision medicine, molecularly targeted agents; optimal design; phase II trials.

Abstract: In the early phase development of molecularly targeted agents (MTAs), a commonly encountered situation is that the MTA is expected to be

more effective for a certain biomarker subgroup, say marker-positive patients, but there is no adequate evidence to show that the MTA does not work for the other subgroup, i.e., marker-negative patients. After establishing that marker-positive patients benefit from the treatment, it is often of great clinical interest to determine whether the treatment benefit extends to marker-negative patients. We propose multi-stage optimal sequential trial (MOST) designs to address this practical issue in the context of phase II clinical trials. The MOST designs evaluate the treatment effect first in marker-positive patients and then in marker-negative patients if needed. The designs are optimal in the sense that they minimize the expected sample size or the maximum sample size under the null that the MTA is futile. We proposed an efficient, accurate optimization algorithm to find the optimal design parameters. One important advantage of the MOST design is that the go/no-go interim decision rules are specified prior to the trial conduct, which makes the design particularly easy to use in practice. A simulation study shows that the MOST designs perform well and are ethically more desirable than the commonly used marker-stratified design. We apply the
MOST design to an endometrial carcinoma trial.

Novel clinical trial design, computation, and implementation for cancer therapy in the precision medicine era

J. Jack Lee, Ph.D.
Department of Biostatistics, University of Texas MD Anderson Cancer Center, USA
Email: jjlee@mdanderson.org

Keywords: Adaptive Design, Adaptive Randomization, Bayesian Update, Platform Design, Predictive Probability, Prognostic and Predictive Biomarkers

Abstract: Rapid advancements in biology demand innovative methods to identify better therapies and the corresponding populations in a timely, efficient, accurate, and cost-effective way. Traditional clinical trials focus on testing a small number of agents in relatively homogeneous populations without the emphasis of tailored treatments for patients enrolled in the trials. With next generation sequencing, no two patients are exactly alike. Great challenges arise to develop agents with unknown prognostic and predictive biomarkers and the desire to provide the most effective, tailored treatment for each patient. In my talk, I will discuss several solutions to meet these challenges including adaptive randomization, predictive probability, multi-arm multi-stage adaptive design, etc. The concept of platform design, Bayesian update, and adaptive learning will also be discussed. Computation aspects for such trials will be discussed. Examples of recent trials will be compared and contrasted. Adaptive biomarker-driven trials can lead to a new paradigm of integrating clinical practice with clinical research to meet the goals of testing treatment efficacy, identifying and validating markers, as well as treating each patient best with precision medicine during the trial.

IS43 EMERGING AREAS IN APPLIED STATISTICS
Session Organizer: Smarajit Bose, Indian Statistical Institute, India
Session chair: Chun-Houh Chen, Academia Sinica, Taiwan
Venue: Seminar Room 1
Time: 19 Dec, 9:00-11:00

Residual based tree and ensemble for clustered binary data

Mousumi Banerjee
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
Email: mousumib@umich.edu

Keywords: clustered data; tree; forest; residuals.

Abstract: Tree-based methods are widely used for classification in health sciences research, where data are often clustered. In this paper, we extend the original classification tree paradigm (CART) (Breiman et al. 1984) to clustered binary outcome setting where covariates are observed both at the cluster- and individual- levels. Using residuals from a null generalized linear mixed model as the outcome, we build a regression tree to partition the covariate space into rectangles. This circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. Class predictions for each terminal node in the final tree are given based on the success probabilities for the specific node. Based on extensive simulations, we compare our residual based classification tree to CART. The methods are illustrated using data from a kidney cancer study and a childhood vaccination study. Finally, to gain accuracy in predictions and address instability in a single tree, we provide extension of our methodology to grow an ensemble of trees or forest.

Bayesian modeling and selection of genetic pathways and genes for cancer

Sounak Chakraborty
University of Missouri, USA

Abstract: Much attention has been given to the development of methods that utilize the large quantity of genetic information available in online databases. Recently a new philosophy emerged which considers the genetic pathways, which contain sets of genes, they have a combined effect on a disease. Under this new idea the goal is to identify the significant genetic pathways and the corresponding influential genes and their combined effect towards different diseases. In this article we propose a Bayesian kernel machine model for right censored survival data which incorporates existing information on pathways and gene networks in the analysis of DNA microarray data. Each pathway is modeled nonparametrically using a reproducing kernel Hilbert space. Mixture priors on the pathway indicator variables and the gene indicator variables are

assigned. This helps us to model both linear and non-linear pathway effects, pinpoint the important pathways along with the active genes within each pathway. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for our model. Simulation studies and a real data analysis, using, van't Veer et al. (2002) breast cancer microarray data, are used to illustrate the proposed method.

Count data methodologies with applications in health sciences

Swarnali Banerjee
Department of Mathematics and Statistics, Old Dominion University, USA
E-mail: sbanerje@odu.edu

Keywords: Count data; Confidence Interval; Purely Sequential; Two-Stage; Fixed-Accuracy; Estimation

Abstract: Count data arising from various areas of statistical ecology may be modelled by a Negative Binomial (NB) Distribution. Existing estimation methodologies for the parameters of an NB distribution were reviewed by Mukhopadhyay and Banerjee (2015). Here, through applications in simulated and a real infestation data set, drawbacks of these methods were highlighted. As a resolution, Mukhopadhyay and Banerjee (2014) introduced a fixed accuracy confidence interval for the NB mean using purely sequential and two stage-procedures.
As such count data are not limited ecology, Banerjee and Mukhopadhyay (2015) generalized the sequential fixed accuracy confidence interval for estimating any positive parameter, discrete or continuous, coming from an arbitrary distribution. We proved that the procedure is asymptotic first- order efficient and asymptotic consistent. Beyond ecology, the procedure finds its applications in other areas like radioactive decay, number of cells affected by some disease, number of cells under genetic mutation, etc. In particular, we gave examples from three specific distributions - Bernoulli(p) distribution for odds-ratio estimation, a Poisson(θ) distribution for mean estimation, and a Normal (θ, θ) distribution for estimation. Large scale simulations results validate the performance of this procedure. We further broaden this general approach to include estimation of a parameter whose parameter space is R. Although this procedure works well under some regularity conditions, it falls through in certain cases such as the proportion parameter for Bernoulli distribution, parameters of a Zero Inated Poisson or Zero Inated Gamma distribution etc. Banerjee and Mukhopadhyay (2015) introduce a new bounded length confidence interval for Bernoulli(p) using both purely sequential and two-stage methodologies. The real data sets analyzed for this paper involved estimating chances of relapse in bone marrow transplant patients (size: small), estimating chances of getting diabetes for Pima Indians (size: moderate) and estimating probability of infestation in a potato beetle data (size: large). Ongoing research involves similar extensions to other distributions.

Robust speaker identification

Smarajit Bose
Indian Statistical Institute, India

Abstract: A novel solution to the speaker identification problem is proposed through minimization of the statistical divergence between the (hypothetical) probability distribution (g) of feature vectors from the test utterance and the probability distributions of the feature vector corresponding to the speaker classes. This approach is made more robust to the presence of outliers, through the use of suitably modified versions of the standard divergence measures. Three such measures were considered – the Likelihood Disparity, the Hellinger distance and the Pearson chi-square distance. The proposed approach was motivated by the observation that, in the case of the Likehood Disparity, when the empirical distribution function is used to estimate g, it becomes equivalent to maximum likelihood classification with Gaussian Mixture Models (GMMs) for speaker classes, a highly effective approach proposed by Reynolds (1995). Significant improvement in classification accuracy is observed under this approach on the benchmark speech corpus NTIMIT and a new bilingual speech corpus NISIS. Moreover, the ubiquitous principal component transformation, by itself and in conjunction with the principle of classifier combination, is found to enhance the performance further.

IS44 COVARIANCE COMPUTATION
Session Organizer and Chair: Sanjay Chaudhuri, National University of Singapore, Singapore
Venue: Seminar Room 2
Time: 17 Dec, 16:00-18:00

Bayesian inference for Gaussian graphical models beyond decomposable graphs

Kshitij Khare
Department of Statistics, University of Florida, USA
E-mail: kdkhare@stat.u.edu

Keywords: Gaussian graphical models, Gibbs sampler, Generalized Bartlett graph, Generalized G-Wishart distribution, Scalable Bayesian inference

Abstract: Bayesian inference for graphical models has received much attention in the literature in recent years. It is well known that when the graph G is decomposable, Bayesian inference is significantly more tractable than in the general non-decomposable setting. Penalized likelihood inference on the other hand has made tremendous gains in the past few years in terms of scalability and tractability. Bayesian inference, however, has not had the same level of success, though a scalable Bayesian approach has its

respective strengths, especially in terms of quantifying uncertainty. To address this gap, we propose a scalable and flexible novel Bayesian approach for estimation and model selection in Gaussian undirected graphical models. We first develop a class of generalized G-Wishart distributions with multiple shape parameters for an arbitrary underlying graph. This class contains the G-Wishart distribution as a special case. We then introduce the class of Generalized Bartlett (GB) graphs, and derive an efficient Gibbs sampling algorithm to obtain posterior draws from generalized G-Wishart distributions corresponding to a GB graph. The class of Generalized Bartlett graphs contains the class of decomposable graphs as a special case, but is substantially larger than the class of decomposable graphs. We proceed to derive theoretical properties of the proposed Gibbs sampler. We then demonstrate that the proposed Gibbs sampler is scalable to significantly higher dimensional problems as compared to using an accept-reject or a Metropolis-Hasting algorithm. Finally, we show the efficacy of the proposed approach on simulated and real data.

Estimation of spectra of high-dimensional time series and its applications

Debashis Paul
Department of Statistics, University of California, Davis, USA
E-mail: debpaul@ucdavis.edu

Keywords: Quadratic forms; random matrix; time series; spectral decomposition.

Abstract: We consider estimating the spectra of the coefficients and autocovariance matrices of a high-dimensional linear time series with simultaneously diagonalizable coefficient matrices. The estimation procedure uses Stieltjes transforms of symmetrized sample autocovariance matrices as input and involves a nonlinear optimization procedure. This procedure also enables us to develop an estimation strategy for a class of quadratic forms involving the inverse of the covariance matrix or the long-term covariance matrix of the observations. Practical performance of the estimators and measures of uncertainty quantification are discussed. This is based on joint work with Haoyang Liu and Alexander Aue.

Matrix-free conditional simulations for spatial Gaussian linear mixed models

Somak Dutta
Iowa State University, USA

Abstract: We develop a matrix-free method for conditional simulations for spatial linear mixed models on regular rectangular lattice. The spatial random fields on the regular lattice are derived by solving fractional Laplacian differenced equations and they provide novel approximations to continuum

Mat\'ern random fields. Our sampling method is exact in contrast to approximate algorithms such as Gibbs sampler, Chebyshev polynomial approximations, and does not require artificially expanding the array as done in circulant embedding. The fundamental principle behind our method consists of constructing a matrix-free square root of the conditional precision matrix and a conjugate gradient recipe to solve system of linear equations involving this conditional precision matrix. The key ingredients are the two-dimensional discrete cosine transformations and a sparse incomplete Cholesky decomposition of the dense conditional precision matrix.
Together they bring down the computational cost of the sampling method to $O(rc(\log (rc))^2)$ and the storage requirement to $O(rc)$ where r and c are the dimensions of the observed array. Furthermore we provide use of the conditional simulation to 1) compute simultaneous exceedance regions and contour sets of the underlying spatial random field, 2) compute the marginal likelihood by Monte Carlo integrating over missing observations, 3) quantify prediction uncertainty of the latent spatial field and 4) sample from the posterior distribution of the spatial field and other parameters under a Bayesian regime. We demonstrate our method on a data on groundwater arsenic concentration in Bangladesh.

IS45 MACHINE LEARNING AND STATISTICAL SIGNAL PROCESSING
Session Organizer: Su-Yun Huang, Academia Sinica, Taiwan
Session Chair: I-Ping Tu, Academia Sinica, Taiwan
Venue: Seminar Room 2
Time: 18 Dec, 10:20-12:20

Tuning parameter selection for low-rank matrix completion methods

Fumitake Sakaori
Department of Mathematics, Chuo University, Japan
E-mail: sakaori@math.chuo-u.ac.jp

Keywords: matrix completion; soft-impute; Information Criterion.

Abstract: Matrix completion methods for low-rank matrix, e.g., Soft-Impute (Mazumder et al., 2010) and Hierarchical Adaptive Soft Impute (Todeschini et al., 2013), have studied and used in many practical situation such as collaborative filtering and analysis of gene expression array. The objective of these methods are to complete an incomplete observed matrix under the assumption that the observed matrix are represented as the sum of a low-rank matrix and a noise matrix. The effectiveness of these methods heavily rely on the choice of tuning parameters, but no attention has paid for it. In this study, we propose some methods of selecting optimal tuning parameters by a cross-validation-based method and some information criteria.

Functional inverse regression in an enlarged dimension reduction space

Ting-Li Chen
Institute of Statistical Science, Academia Sinica, Taiwan
E-mail: tlchen@stat.sinica.edu.tw

Keywords: Functional inverse regression; functional dimension reduction; functional linearity condition; sliced inverse regression.

Abstract: We consider an enlarged dimension reduction space in functional inverse regression. Our operator and functional analysis based approach facilitates a compact and rigorous formulation of the functional inverse regression problem. It also enables us to expand the possible space where the dimension reduction functions belong. Our formulation provides a unified framework so that the classical notions, such as covariance standardization, Mahalanobis distance, SIR and linear discriminant analysis, can be naturally and smoothly carried out in our enlarged space. This enlarged dimension reduction space also links to the linear discriminant space of Gaussian measures on a separable Hilbert space.

GenSVM: multiclass support vector machines

Patrick J.F. Groenen
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands

Abstract: For binary classification problems, support vector machines (SVMs) have become increasingly popular over the last ten years. However, in the case that more than two classes need to be predicted often a series of binary SVMs are performed (one-versus-all or between all pairs of classes, one-versus-one). A disadvantage of such methods is that they are heuristics that do not simultaneously estimate all parameters in a single model.
In this presentation, we discuss a new multiclass SVM loss function (GenSVM) that is based on a geometric representation of each class by a vertex of a simplex in $K-1$ dimensional space. As with the binary SVM, an object that is predicted to be nearest to its class receives a zero error and if the object is closer to another class the error consists of a function of the distance to the zero-error region. The present approach is flexible in the hinge function that is used for calculating the error. It builds on the Huberized hinge errors that have as special cases the linear and quadratic hinges. It is also flexible in how these errors are added: we propose to use the L_p norm of the Huberized hinge error. This general loss function has some existing multiclass SVM loss functions as special cases. Provide a majorization algorithm that minimizes GenSVM. Numerical comparisons show that for medium sized problems GenSVM compares with the best approaches.

Variable selection and predictive performance for correlated biomarkers using regularized regression approaches

Wei-Ting Hwang, PhD
University of Pennsylvania, USA

Abstract: The goal in many biomedical studies is to identify biomarkers that predict patient phenotypes such as disease status or treatment response. Evaluation of biomarker signature composition and its predictive performance are critical. Recent regularization approaches such as LASSO, Elastic-net or their extensions are increasing popular as a tool for variable selection in identifying an optimal subset of biomarkers from a group of high-dimensional candidate biomarkers. However, many of those variable selection methods tend to give unstable results in particular if the correlations between candidate biomarkers are high. In this project, we conduct a series of simulation study to understand the impact of the correlation among predictors and other factors (e.g., sample size, number of candidate biomarkers, presence of confounders, etc.) on signature composition and model performance. We also compare a more efficient nested cross-validation process to the standard cross-validation in selecting tuning parameters values under various scenarios. The presented work will focus on binary disease status as response variable and continuous biomarker candidates as predictors through logistic regression. Application on a real world data for mesothelioma biomarker discovery will be presented.

IS46 ANALYSIS AND APPLICATIONS OF OMICS DATA
Session Organizer and Chair: Hsuan-Yu Chen, Academia Sinica, Taiwan
Venue: Stephen Riady Global Learning Room
Time: 18 Dec, 15:50-17:50

Identifying driver mutations in cancer through constrained genotype-phenotype association mining

Niranjan NAGARAJAN
Genome Institute of Singapore, A*STAR，Singapore

Abstract: Extensive and multi-dimensional data sets generated from recent cancer omics profiling projects have presented new challenges and opportunities for unraveling the complexity of cancer genome landscapes. In particular, distinguishing the unique complement of genes that drive tumorigenesis in each patient from a sea of passenger mutations is necessary for translating the full benefit of cancer genome sequencing into the clinic. We address this need by presenting a data integration framework (OncoIMPACT) to nominate patient-specific driver genes based on their phenotypic impact. Extensive in silico and in vitro validation helped establish OncoIMPACT's robustness, improved precision over competing approaches and verifiable patient and cell line specific predictions. In particular, we computationally

predicted and experimentally validated the gene TRIM24 as a putative novel amplified driver in a melanoma patient. Applying OncoIMPACT to more than 1000 tumor samples, we generated patient-specific driver gene lists in five different cancer types to identify modes of synergistic action. We also provide the first demonstration that computationally derived driver mutation signatures can be overall superior to single gene and gene expression based signatures in enabling patient stratification and prognostication. Source code and executables for OncoIMPACT are freely available from http://sourceforge.net/projects/oncoimpact.

Integration of string and de Bruijn graphs for genome assembly

Yao-Ting Huang
Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan

Abstract: String and de Bruijn graphs are two graph models used by most genome assemblers. At present, none of the existing assemblers clearly outperforms the others across all datasets. We found that although a string graph can make use of entire reads for resolving repeats, de Bruijn graphs can naturally assemble through regions that are error-prone due to sequencing bias. We developed a generalized assembler called StriDe that has advantages of both string and de Bruijn graphs. First, the reads are decomposed adaptively only in error-prone regions. Second, each paired-end read is extended into a long read directly using an FM-index. The decomposed and extended reads are used to build a generalized assembly graph. In addition, several essential components of an assembler were designed or improved. The resulting assembler was fully parallelized, tested, and compared with state-of-the-art assemblers using benchmark datasets. The results indicate that contiguity of StriDe is comparable with top assemblers on both short-read and long-read datasets, and the assembly accuracy is high in comparison with the others.

Integrative genomic analysis of lung cancer cell lines

Yi-Chiung Hsu
Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Lung cancer is the leading cause of cancer mortality worldwide. Genetic analyses and gene expression profiling of human lung tumors identified several aberrant signaling pathways involved in the lung cancers. Genetic alterations in cancers have been linked with response to targeted therapeutics and tumor metastasis on activated oncogenic signaling pathways. We collected 44 lung cancer cell lines with variety of histology patterns, EGFR status and invasion ability. We generated several genomics data such as DNA copy number variation and Next-Generation-Sequencing Cancer Panel. By correlating with the lung cancer cell genomic profiles, genetic predictors of invasion ability were available; but they did not connect well with clinical

outcomes. Because metastasis is another essential factor of cancer mortality, we conducted an invasion assay to obtain the phenotype information of lung cancer cell lines. After a series of statistical analysis, we obtained a set of invasion-associated genes. We validated the signature's performance with early stage lung cancer cohort. Our data light on the complex genetic involvement in the molecular mechanisms of tumor invasion and growth inhibition.

Estimation of imputed genome variations and the risk for hepatocellular carcinoma among chronic hepatitis c patients

Mei-Hsuan Lee
Institute of Clinical Medicine, National Yang-Ming University, Taiwan

Abstract: Hepatitis C virus (HCV) infects more than 185 million individuals worldwide. Approximately 80% of acute infection may progress to chronic infection, 20% of chronic hepatitis C patients may develop liver cirrhosis within 25 years, and 25% of cirrhosis patients may occur hepatocellular carcinoma (HCC). Host genetic susceptibility may be associated with the risk for the occurrence of HCC among patients infected with HCV. We aimed to discover the genomic variations associated with HCC risk through the genome-wide association study and imputation analysis. We included 472 HCC cases and 806 unaffected controls in this study. All of the study subjects were adults seropositive for antibodies against HCV and seronegative for HBsAg. The demographic characteristics, serum markers including liver functions and viral factors were collected and evaluated as well. High quality human genomic DNA was extracted from blood sample of each subject in order to perform the genotyping by AxiomTM Genome-Wide CHB Array. Then, the imputation algorithm was applied to obtain whole genome variations in patients based on genotyping data of single nucleotide polymorphism (SNP) arrays. The Han Chinese population in 1000 genome project was used as a reference for imputation analysis. The logistic regression models were utilized to evaluate the magnitude of the associations between HCC and genotype based on four different genetic models (allelic, dominant, additive and recessive). After imputation analysis, a total of 36,175,343 SNPs were obtained for each subject. There were 765, 134, 612, and 855 SNPs significantly associated with HCC based on additive, recessive, dominant, and allelic genetic models, in correspondingly. There were 18 non-synonymous SNPs were found to be significantly associated with HCC in various genetic models. Interestingly, there were seven non-synonymous SNPs clustered in the human leukocyte antigen complex region, which is a relevant region associated with human immunological responses. The findings suggested that the genetic variants were important for the associations of HCC, which provided information to identify high risk population among patients with HCV infection.

IFCS CLUSTER ANALYSIS AND MULTIDIMENSIONAL SCALING IN ANALYSIS OF MARKETING DATA
Session Organizer and Chair: Akinori Okada, Tama University, Japan
Venue: Stephen Riady Global Centre Learning Room
Time: 17 Dec, 10:40-12:40

Classification and prediction of topics on social media considering temporal variation

Atsuho Nakayama
Graduate School of Social Sciences, Tokyo Metropolitan University, Japan
E-mail: atsuho@tmu.ac.jp

Keywords: Complementary Similarity Measure, Multidimensional Scaling, Non-negative Matrix Factorization, Over-lapping Clustering, Social Listening, Text Analysis

Abstract: This study examined word clustering in social media, focusing on new products. We collected Twitter entries about new products, based on specific sentiment or interest expressions. To identify market trends, the analysis of consumer tweet data has received much attention. In this study, we examined temporal variation in topics on new products by classifying words into clusters, based on the co-occurrence of words in Twitter entries. We selected keywords representing meaningful topics. To construct appropriate words, we used a complementary similarity measure, which is a classification method widely applied to characters on a time-series basis. The selected entry-by-word matrix had high dimensions, so it was necessary to analyze for dimensionality reduction. We classified the words extracted from Twitter data using non-negative matrix factorization for a dimensionality reduction model. Then, we proposed a visualization method for text classification to interpret the results, using a multidimensional scaling model. Finally, we detected some topics about new products by classifying words into clusters based on the co-occurrence of words in Twitter entries, and weekly tendencies of topics about new products by classifying words into clusters based on the co-occurrence of words in Twitter entries.

Additive clustering for asymmetric similarity matrices and its application for market segmentation

Tadashi Imaizumi
School of Management & Information Sciences, Tama University, Japan
E-mail: imaizumi@tama.ac.jp

Keywords: Over-lapping Cluster, Common Features, Distinctive Features, NMF, Optimization

Abstract: The product positioning and market segmentation are important studies for predictive analysis in marketing science. Multidimensional scaling

and cluster analysis are typical models and methods to extract information from similarity matrices. These models and methods are appropriate when the supplied similarity matrices are symmetric. However, a brand switching matrix, for example, is asymmetric. So, the models and methods for asymmetric similarity matrices are needed to be studied. The following three models are general models to analyze asymmetric similarity matrices, 1) Distance model, 2) Feature Model, and 3) Similarity Choice model. Okada and Imaizumi proposed several model and non-metric methods which based on distance model. Olszewski also proposed K-Means Clustering for asymmetric data. Each of them has some difficulty as the number of objects is increased. And a new additive clustering, one of Feature Model, will be proposed for asymmetric similarity matrices. The proposed model explains similarity matrices by the common features and the distinctive features of a subset. The optimization criterion based on L1 norm will be proposed. An initial allocation of objects is done by using a Nonnegative Matrix Factorization. An application to real data sets are also shown. Results show asymmetric similarity matrices were accounted by the proposed model and method.

Evaluating offensive and defensive power in brand switching by asymmetric multidimensional scaling

Akinori Okada
Research Institute, Tama University, Tokyo, Japan
E-mail: okada@rikkyo.ac.jp

Keywords: Asymmetry, Brand switching, Inward tendency, Multidimensional scaling, Outward tendency.

Abstract: Brand switching is asymmetric and thus is to be analysed by asymmetric multidimensional scaling (Borg & Groenen, 2005, Ch. 23). When there are brand switching data for periods i, i+1, and i+2, we have two brand switching matrices; the first corresponds to periods i to i+1, and the second corresponds to periods i+1 to i+2. The asymmetric multidimensional scaling (Okada and Tsurumi (2012), based on the SVD (singular value decomposition), of each matrix gives the outward tendency (offensive power in brand switching) and the inward tendency (defensive power in brand switching) of each brand. The external analysis of the second matrix using the inward/outward tendency of the first matrix gives the outward/inward tendency of the second matrix. The evolution of offensive/defensive power is evaluated by comparing the outward/inward tendency of the second matrix derived by the original (not external) analysis with that derived by the external analysis shows.

ISBIS INNOVATIVE STATISTICAL METHODS IN BUSINESS AND INDUSTRY
Session Organizer and Chair: Yuli Hong, Virginia Tech, USA
Venue: LT52
Time: 18 Dec, 13:30-15:30

Analysis of location and variation transmission in the multi-stage manufacturing process and its application to process monitoring

Changsoon Park
Department of Statistics, Chung-Ang University, Seoul, South Korea

Abstract: In the multi-stage manufacturing process, it is important to know how the effects of upstream stages are added and transmitted to quality characteristics (QCs) across the stages of the process. Variation transmission problem is critical in quality management and has been a major issue in the multi-stage production process. On the other hand, the location transmission can be easily adjusted when its amount can be measured, but becomes critical also when it cannot be measured directly. This research discusses both the location and variation transmission problems for cases where QCs of stages are not directly measurable. When the QCs cannot be measured due to metrology problems and/or measurement errors, the structure of the QCs at each stage is expressed as a structural equation model (SEM). In the SEM, the QCs are regarded as the latent variable which is expressed as a linear regression function of the previous QC estimates and the current state input variables. Instead of the QC measurements themselves, their indicators are measured to represent the QCs. One powerful method for estimating the QCs as well as estimating the regression parameters is to use the Bayesian SEM approach where the QCs are treated as missing values. The QCs and the structure parameters are estimated by the Bayesian MCMC method using the Gibbs sampler. The monitoring of the process quality can be done in two ways: Direct monitoring of the QC estimates and profile monitoring of the regression parameters. Directing monitoring is conveniently used for out-of-control signal, and the profile monitoring can be used in detecting the source of the variation.

Disentangling and Assessing Uncertainties in Multi-period Corporate Default Risk Predictions

Yili Hong
Department of Statistics, Virginia Tech, Blacksburg VA 24061

Abstract: Measuring credit risks for individual companies, industrial segments, and market systems is fundamentally and broadly important in economics and finance. For such a purpose, various quantitative methods have been developed to predictively assess the probabilities of companies going default in future. However, as a more difficult and crucial problem, evaluating the uncertainties associated with the default predictions remains little explored. In

this paper, for the first time in the scenario of default predictions, we develop a procedure for quantifying the level of associated uncertainties by carefully disentangling multiple contributing sources. Our framework effectively incorporates broad information from historical default data, financial records, and macroeconomic conditions by a) characterizing the default mechanism, and b) capturing the future dynamics of various features contributing to the default mechanism. Our development of the framework overcomes major challenges in this tremendously large scale statistical inference problem and makes it practically feasible by using parsimonious models, innovative methods, and modern computational facilities. By appropriately predicting the market-wide total number of defaults and assessing the associated uncertainties, our method can effectively evaluate the aggregated market credit risk level. Upon analyzing a US market data set with our method, we demonstrate that the level of uncertainties associated with default risk assessments is indeed substantial. More importantly and informatively, we also find that the level of uncertainties associated with the default risk predictions is correlated with the level of default risks, indicating potential for benefiting practical applications including improving the accuracy of default risk assessments.
This is a joint work with Miao Yuan, Cheng Yong Tang, and Jian Yang.

Statistical modeling and analysis of competing risk data from repairable systems

Anupap Somboonsavatdee
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand

Abstract: The focus of this talk is on failure history of repairable systems for which the relevant data comprise successive event times for a recurrent phenomenon along with an event-count indicator. Such data commonly occur both in industrial and biomedical contexts. In an industrial setting, typically a single expensive prototype of a complex system is observed until failure, followed by corrections or design changes, and subsequent retesting. In the biomedical context, however, it is more common to encounter data on multiple subjects with only a few recurrences per subject. In this talk we shall describe the findings from a study of failure data both from a single repairable system and multiple repairable systems to multiple failure modes, a framework traditionally dubbed as competing risks. We adopt a parametric premise and discuss the results under the Power Law Process model that has found considerable attention in describing recurrent hardware failures of complex mechanical systems. Some interesting and non-standard asymptotic results ensue in this context that will be discussed in detail. We will report findings from an extensive simulation study that supplements the theoretical findings. The methodology will be illustrated on recurrent failure data obtained from a warranty claim database for a fleet of automobiles.

Monitoring the results of cardiac surgery based on 3 or more outcomes by general variable life-adjusted display

Gan Fah Fatt
Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, Singapore.

Abstract: The variable life-adjusted display (VLAD) was first proposed as a simple graphical display for monitoring the results of cardiac surgery adjusted by the risks of patients. The VLAD is now popularly used all over the world. The VLAD assumes binary outcomes: death within 30 days of an operation or survival beyond 30 days. This naive classification of outcomes means that for example, a fully recovered patient is considered the same outcome as another patient who is bedridden for life, and this result is in a great loss of information. We develop a general VLAD that is based on three or more outcomes and this more refined procedure will reflect more accurately and fairly the performance of a surgeon.

Contributed Session

CS01 RECENT DEVELOPMENT OF MULTIVARIATE ANALYSIS FOR HIGH-DIMENSIONAL DATA
Session Organizer: Hirofumi Wakaki and Hirokazu Yanaghihara, Hiroshima University, Japan
Session Chair: Mariko Yamamura, Hiroshima University, Japan
Venue: Seminar Room 3
Time: 17 Dec, 10:40-12:40

Consistent information criterion in normal multivariate linear regression models even under high-dimensionality

Hirokazu Yanagihara
Department of Mathematics, Graduate School of Science, Hiroshima University.
E-mail: yanagi@math.sci.hiroshima-u.ac.jp

Keywords: Consistency; High-dimensional asymptotic framework; Information criterion; Variable selection.

Abstract: There are consist information criteria for selecting variables in normal multivariate linear regression model when the sample size goes to infinity. When the sample size and the dimension of response variables vector go to infinity simultaneously, whether an information criterion is consistent or not strongly depends on a speed of the divergence of a noncentrality matrix. At present, there is no information criterion which is always consistent. Hence, in this paper, new consistent information criterion even under any situations is proposed by adding a constant penalty term, which depends only the sample size, the dimension and the number of explanatory variables, to a negative twofold maximum log-likelihood. Although this criterion is proposed for normal multivariate linear regression models, there is a possibility that our criterion can be applied to other information criteria consisting of a difference between log-determinants of two Wishart matrices.

Kernel function used in the optimum nonlinear discriminant analysis

Takio Kurita
Department of Information Engineering, Hiroshima University.
E-mail: tkurita@hiroshima-u.ac.jp

Keywords: Fisher discriminant analysis, kernel learning, discriminant kernel

Abstract: Recently the kernel learning such as kernel support vector machine or the kernel discriminant analysis has been successfully applied in many applications. However, the kernel function is usually defined a priori and it is not known what the optimum kernel function for classification is. Also the class information is not usually used to define the kernel function. On the other hand, Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities between the feature vectors and

classes in 1975. The ONDA is closely related with Bayesian decision theory. In this talk we derive the optimal kernel function in terms of the discriminant criterion by investigating the optimum discriminant mapping constructed by the ONDA. It is given as an inner product of the vectors defined from the normalized Bayesian posterior probabilities. For real applications, we also introduce a family of discriminant kernels by changing the estimation method of the Bayesian posterior probabilities. The effectiveness of the proposed discriminant kernels is verified through the experiments using UCI ML repository datasets.

Regularization parameter selection for the LASSO in generalized linear models

Yoshiyuki Ninomiya
Institute of Mathematics for Industry, Kyushu University

Keywords: AIC; convexity lemma; non-concave penalty; statistical asymptotic theory; tuning parameter; variable selection

Abstract: The LASSO contains a regularization parameter that determines the result, and several information criteria have been proposed for its selection. While any of them would assure consistency in model selection, we have no appropriate rule to choose between the different possible criteria. On the other hand, a finite correction to the AIC has been provided in a Gaussian linear regression setting. The finite correction is theoretically assured from the viewpoint not of the consistency but of minimizing the prediction error, and it does not have the above-mentioned difficulty in the choice. In general, however, the finite correction cannot be obtained in the case of generalized linear models, and so we derive a criterion from the original definition of the AIC, that is, an asymptotically unbiased estimator of the Kullback-Leibler divergence. Our criterion can be easily obtained and requires fewer computational tasks than does cross-validation, but its performance is almost the same as or superior to that of cross-validation. Moreover, our criterion is generalized for a class of other regularization methods. This is joint work with Shuichi Kawano and Yuta Umezu.

Clustering-based models for high-dimensional data

Mika Sato-Ilic
Faculty of Engineering, Information and Systems, University of Tsukuba
E-mail: mika@risk.tsukuba.ac.jp

Keywords: clustering, high-dimensional data, scale-based models

Abstract: Recent advances in the area of information science have enabled the collection of high-dimension and complex data in vast amounts. Model-based clustering has been tasked with the increasingly significant mission of dealing with such data and the main issue of the model-based clustering is an

assumption of a model to the data and by fitting the model to data, an adjusted partition will be estimated. Although this approach has the benefit of obtaining a clear solution as the result of the partition based on mathematical theory, we cannot avoid the risk the previously assumed model might not adjust to the latent classification structure of the data. Therefore, we propose a framework called clustering-based models in which we exploit the obtained clustering result as a scale of latent structure of the data and apply it to the observed high-dimensional data, and then apply the modified data to a model in order to obtain a more accurate result. In this talk, several clustering-based models within this framework along with several applications will be introduced.

CS02 ROBUST METHODS IN ANALYZING BIG DATA

Session Organizer and Chair: Erniel B. Barrios, University of the Philippines Diliman, Philippines
Venue: Seminar Room 4
Time: 17 Dec, 10:40-12:40

Estimation of multiple time series with volatility

Mark Louie Ramos
Department of Mathematics and Physics, University of Santo Tomas.
E-mail: mframos@mnl.ust.edu.ph

Keywords: Multiple time series, volatility, modeling, bootstrap

Abstract: Volatility episodes in a multiple time series exert influence on the parameter estimates that will significantly reduce model fit for the non-volatile parts. Assuming a multiple time series with common autoregressive parameter, a combination of block bootstrap and forward search algorithm embedded into a backfitting algorithm was used to generate robust estimates when temporary volatility is present in the data. Simulation studies exhibited robustness of the estimates from the hybrid algorithm. Furthermore, predictive ability of the fitted model is high during the non-volatile periods of the time series. While predictive ability deteriorates when the time series are very short or when they are nearly non-stationary, it still provides better estimates than some common modeling strategies.

Robust simultaneous confidence interval estimation of principal component loadings

Martin Augustine B. Borlongan
School of Statistics, University of the Philippines - Diliman
E-mail: mbborlongan@gmail.com

Keywords: principal components analysis, bootstrap, forward search algorithm, outliers

Abstract: Asymptotic results are available for confidence interval (CI) estimation concerning the eigenvectors of the covariance matrix. Bootstrap methods were introduced to relax the distributional assumptions imposed on the data. As Principal Components Analysis (PCA) works on the covariance matrix which is sensitive to the presence of outliers, such observations cause perturbations to the eigenvectors, to the principal components and to subsequent inference on them. The proposed method leverages on the bootstrap incorporated within a fully-automated forward search to come up with robust CI estimates for the elements of the eigenvectors of the correlation matrix in the presence of outliers. The coverage probability of the proposed bootstrap simultaneous CI was compared to the coverage probability of the asymptotic confidence region computed from unprocessed data and from subsets of the data with Minimum Covariance Determinant (MCD) through a simulation study under several scenarios. The proposed method was shown to have more stable coverage probabilities for datasets with or without outliers across several sample sizes as compared to the two approaches based on asymptotic results.

Robust estimation of a dynamic spatio-temporal model with structural change

Stephen Jun Villejo
University of the Philippines Diliman, Philippines

Keywords: spatio-temporal model, backfitting algorithm, bootstrap, forward search, dynamic model

Abstract: We postulate a dynamic spatio-temporal model assuming constant covariate effect but with varying spatial effect over time and varying temporal effect across locations. We proposed a backfitting algorithm embedded with forward search algorithm and bootstrap to generate robust estimates of the parameters in the event of a temporary structural change. A simulation study is designed to account for various scenarios. The mean absolute prediction errors (MAPE) under the proposed algorithm are smaller than with ordinary linear model. This is especially true when there are more spatial units than time points. The proposed algorithm also produced lower relative bias and standard errors especially for the spatial parameter estimates. The relative bias declines when there are more neighborhoods. Predictive ability of the models deteriorates when structural change happened in the more recent periods.

CS03 COMPLEX MULTIVARIATE MODELS I
Session Chair: Daniel Paulin, National University of Singapore, Singapore
Venue: Seminar Room 3
Time: 17 Dec, 13:40-15:40

Discriminant analysis based on vine copulas

Yuki Yasuda and Akio Suzukawa
Graduate School of Economics and Business Administration, Hokkaido University. E-mail: divbe@eis.hokudai.ac.jp

Keywords: Classification; Bayes rule; Non-Gaussian.

Abstract: In this paper, we consider classification problem of two populations. One or more new observations are classified into one of the populations based on the measured characteristics. When continuous distributions are assumed for the populations, Bayes discriminant function is given by the log likelihood ratio of the two density functions. If two populations are multivariate Gaussian with common covariance matrices, the discriminant function is linear, and it is called Fisher's linear discriminant function.
Copula is a function that combines multivariate joint distribution and it marginals. When we would like to model multivariate joint distribution, by using copula, we can separately consider only marginals and dependence structure of random variables.
There are two types of multivariate non-Gaussianity. One is marginal non-Gaussianity. Another is copula non-Gaussianity. Even if copula of a continuous multivariate distribution is Gaussian, the joint distribution is non-Gaussian when one of its marginals is non-Gaussian. On the other hand, even if all marginals are Gaussian, joint distribution is non-Gaussian for non-Gaussian copula. How to deal with the marginal non-Gaussianity is not so difficult since marginals are univariate. We can transform all marginals to Gaussian distributions. In this paper, Bayes discriminant functions are derived under copula non-Gaussianity. We use vine copula as the non-Gaussian copula. Multivariate vine copula is hierarchically constructed by using bivariate copulas. Multivariate joint distribution can be exibility modelled by it.
When parameters of the vine copula are unknown, they are estimated by ML method or IFM (inference function for margin) method. We present some properties of the discriminant functions and misclassification rate under the vine copula models.

Generalized Archimedean copulas

Akio Suzukawa
Graduate School of Economics and Business Administration Hokkaido University.
E-mail: suzukawa@econ.hokudai.ac.jp

Keywords: Multivariate lifetime; Shared frailty; Odds ratio.

Abstract: Multivariate survival data occur in many areas, including medicine, biology, engineering, and economics. For multivariate survival data, the standard univariate methods can be applied marginally. However, dependence structures of multivariate lifetimes are ignored by this approach. A useful approach for dependence modeling is a copula approach. In modeling of multivariate lifetimes, copulas appear early on, in Clayton (1978), Hougaard (1986a，1986b), Marshalland Olkin (1988), Heckman and Honore (1989) and Oakes (1989).
Many of models for multivariate lifetimes are based on the concept of shared frailty. The shared frailty models have been discussed in a lot of textbooks of multivariate survival analysis, e.g. Andersen et al. (1993), Hougard (2000), Aalen (2008) and so on. The shared frailty models are closely related to the Archimedean copula introduced by Kimberling (1974). The Archimedean copula has been treated in many textbooks of copula theory，e. g. Nelsen (2006), Joe (2014).
Several generalizations of the Archimedean copula have been proposed. Non-exchangeable generalizations of the Archimedean copula were proposed by McNeil et al. (2005) and Mesiar (2013). Bivariate Archimax copula introduced by Caperaa (2000) and its multivariate extension by Charpentier et al. (2014) are also generalizations of the Archimedean copula. McNeil and Neslehova (2010) extended the Archimedean copula to the Liouville copula. In this paper, we propose a generalization of the Archimedean copula based on the generalized odds ratio introduced by Dabrowska and Doksum (1988). The generalized Archimedean copula is non-exchangeable. Based on it, multivariate lifetime distributions can be flexibly modeled. A sampling algorithm from the generalized Archimedean copula is obtained. Estimation procedures based on multivariate lifetime data are formulated.

Identifiability of Gaussian DAG models with one latent source

Hisayuki Hara
Niigata University, Japan

Keywords: factor analysis; structural equation; gr\"obner basis; tetrad.

Abstract: We study parameter identifiability of directed Gaussian graphical models with one latent variable. In the scenario we consider, the latent variable is a confounder that forms a source node of the graph and is a parent to all other nodes, which correspond to the observed variables. We give a graphical condition that is sufficient for the Jacobian matrix of the parametrization map to be full rank, which entails that the parametrization is generically finite-to-one, a fact that is sometimes also referred to as local identifiability. We also derive a graphical condition that is necessary for such identifiability. Finally, we give a condition under which generic parameter identifiability can be determined from identifiability of a model associated with

a subgraph. This is a joint work with Mathias Drton and Dennis Leung at University of Washington.

A new class of copulas involved geometric distribution: estimation and applications

Kong-Sheng Zhang
Department of Mathematics, Southeast University, China
E-mail: zks155@163.com

Keywords: Copula; geometric distribution; maximum likelihood estimation; interior-point penalty function method

Abstract: Copula is becoming a popular tool for modelling the dependence structure among multiple variables. Commonly used copulas are Gaussian, t and Gumbel copulas. To further generalize these copulas, a new class of copulas, referred to as geometric copulas, is introduced by adding geometric distribution into the existing copulas. The interior-point penalty function algorithm is proposed to obtain maximum likelihood estimation of the parameters of geometric copulas. Simulation studies are carried out to evaluate the efficiency of the proposed method. The proposed estimation method is illustrated with workers' compensation insurance data and exchange rate series data.

CS04 MODELS FOR COMPLEX BIOLOGICAL DATA
Session Chair: Sanjay Chaudhuri, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 17 Dec, 13:40-15:40

Using robust variance estimation in mixed models: a review

A.A.Sunethra, M.R. Sooriyarachchi
University of Colombo, Sri Lanka
Email: sunethra@stat.cmb.ac.lk, roshinis@hotmail.com

Keywords: Correlated Data, Mixed Models, Robust Variance, Random Effects, Sandwich Variance Estimation

Abstract**:** Presence of Clusters/ sub-groups within datasets is a common phenomenon in statistical data analysis. Examples include repeated measures data, longitudinal data, hierarchical data, and etc. The shared feature in such datasets is that observations within a group are related / similar to each other. Data of this kind is termed as correlated data or non-independent data. When analyzing such data, the methods of analysis should

not rely on the assumption of independence which is a dominant assumption in statistics.

Robust Variance Estimation which is often nicknamed as Sandwich Variance Estimation (SVE) is a method of variance estimation initially proposed by Peter J Huber in 1967 to correct the estimation of standard errors of miss-specified models, i.e. in models that are being fitted incorrectly. These miss-specifications/errors may be due to various reasons such as incorrect distributional assumptions, assuming linear relationships for non-linear data, assuming independency for correlated data and etc. This methods gained more popularity with its derivation in linear regression by H. White in 1980 where he demonstrate its usage for independent, heteroscedastic errors in linear regression models where the miss-specifications was not due to independence but due to errors being heteroscedastic. In contrast, with correlated data modeling, SVE requires to cater for heteroscedatsic, non-independent data. Hence, SVE is being a method for adjusting the standard errors of model parameters; it had been extensively used in correlated data analysis for obtaining standard errors that are adjusted to the correlation of the data where the adjustment made by SVE doesn't rely on the model being fitted to the data. In olden days, when statistical models were not developed for correlated data, models assuming independence were fitted for non-independent data and the model standard errors were adjusted by using SVE. The literature had emphasized that SVE has provided improved inferential results in correlated data analysis in the absence of statistical models for correlated by improving the functionality of independency assumed models fitted for correlated data. In addition to the classical SVE, various adjustments for the classical SVE had been developed for various data scenarios such as small sample data, data with auto-correlation and etc.

Lately, specialized statistical models were developed for correlated data such as Mixed Models and Generalized Linear Mixed Models (GLMMs). Since these models are defined for correlated data, the model parameter estimates and standard error estimates are resultant to the correlation exist in the data. Therefore, the necessity of SVE in such models was at argument by authors in the literature. Mixed models are defined in such a way that clusters/groups that impose correlation to the data is being introduced to the model as random effects that follow a particular statistical distribution (Gaussian, Gamma, t-distribution) where the linear predictor of mixed models consists of a component that represent the grouping/clustering in the data. More over the literature consists of few authors that had demonstrated probable miss-specifications of Mixed Models despite they are defined for correlated data. These miss-specifications are mainly due to the disparity between the correlation structure of the data and the way the random effects are defined in Mixed Models. Upon the identification of miss-specifications of such hybrid models, adoption of SVE in GLMMs becomes remedial since SVE is meant for improving miss-specified models. Though SVE was initially proposed for correcting the standard errors of maximum likelihood estimates, it can be used for parameter estimation methods which obtain parameter estimates by equating the estimation function to zero which doesn't necessarily be a derivative of a log-likelihood. Thus, SVE are feasible with Mixed Models hich mostly accommodate pseudo likelihood methods in parameter estimation.

Researches or studies which have looked at the use of SVE particularly in Mixed Models or GLMMs are very few where a research comparing the use of SVE in GLMMs for analyzing two actual datasets was found which showed up evidence for SVE is being capable of correctly estimating the variance of the fixed effects parameters of GLMMs even when random effects are miss-specified. Researches that had highlighted miss-specifications of GLMMs mainly had exposed the errors of random effects definition of GLMMs not being able to represent precisely the correlation structure present in the data. Therefore, modifying random effect definition can be considered as direct solution for this issue whereas SVE serves indirectly by improving the standard error estimation of Mixed Models. Since SVE can improve the estimation of model standard errors, it improves model adequacy tests and other hypothesis test associated with Mixed Models. Simulating correlated data scenarios with probable miss-specifications in par with Mixed Models' random effect definition and then analyzing those data using suitable Mixed Models while using SVE can be used for evaluating feasibility of using SVE in Mixed Models. Further, the sample size and the level of the correlation present in the data could also impact on the performance of Mixed Models. The development of various adjustments for the classical SVE had mainly taken place for coping up with various correlation structures and for dealing with small sample size where SVE was earlier considered as an asymptotical method which works well for large sample sizes. Therefore, the choice of the SVE adopted should carefully be made with respect to the correlation structure present in the data and with respect to the sample size of the data. It was identified through simulation that Mixed Models with SVE assist on enhancing its functionality than used with standard method of variance estimation while at small sample sizes enhancements can be achieved by using small sample adjusted SVEs. In summary, it can be emphasized though SVE is being a method of variance estimation developed nearly about half a century before, its applicability still resides even with hybrid statistical models like Mixed Model or GLMMs which are very recently developed statistical modeling approaches for correlated data.

Influence analysis of area under ROC curve

Bo-Shiang Ke
National Chiao Tung University, Taiwan

Keywords: AUC, influence function, local influence, cumulative lift chart

This work is collaborated with Yuan-chin Ivan Chang at Institute of Statistical Science, Academia Sinica.

Abstract: Supervised learning is a major issue in statistical learning, especially binary classification problems. Enormous techniques are created to deal with them. In order to quantify the performances of classifiers, an objective performance criterion is indispensable. Area under ROC curve (AUC) is a popular performance measure due to its elegant interpretation; however, potential influential observations may alter its conclusion. To this end, we first

utilize the cumulative lift chart to visualize the existence of the potential influential observation and its approximate location. Then we adopt the theoretical approaches such as the influence functions of Hampel (1974) and the local influence of Cook (1986) and Wu and Luo (1993) for AUC estimation. Based on theoretical results, we are able to distinguish potential influential observations from others via the inner fences proposed by Tukey (1977). Owing to the lack of similar detectable techniques in binary classification assessments, these graphical and theoretical approaches will enhance more insights for data analysts.

Melanoma cell colony expansion parameters revealed by approximate Bayesian computation

Brenda N Vo
Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia.
Email: n1.vo@qut.edu.au

Keywords: Approximate Bayesian computation; Sequential Monte Carlo; Cell diffusivity and proliferation; Cell-to-cell adhesion; Random walk model; Melanoma cell colonies.

Abstract: In vitro studies and mathematical models are now being widely used to study the underlying mechanisms driving the expansion of cell colonies. This can improve our understanding of cancer formation and progression. Although much progress has been made in terms of developing and analysing mathematical models, far less progress has been made in terms of understanding how to estimate model parameters using experimental in vitro image-based data. To address this issue, a new approximate Bayesian computation (ABC) algorithm is proposed to estimate key parameters governing the expansion of melanoma cell (MM127) colonies, including cell diffusivity, D, cell proliferation rate, λ, and cell-to-cell adhesion, q, in two experimental scenarios, namely with and without a chemical treatment to suppress cell proliferation. Even when little prior biological knowledge about the parameters is assumed, all parameters are precisely inferred with a small posterior coefficient of variation. The ABC analyses reveal that D and q depend on the experimental elapsed time, whereas λ does not. Furthermore, we found that q also depends on the initial cell density, whereas D and λ do not. The ABC approach also enables information from the two experiments to be combined, resulting in greater precision for all estimates of D and λ.

Elucidation of the aortic aneurysm mechanism by cooperation of mathematical science and Cardiovascular Surgery

Saki Goto
Graduate School of Environmental and Life Science, Okayama University.

Abstract: Aortic aneurysm is a disease requiring the treatment, because there are various risks in surgery, there is a need for criteria for surgery for each case. If not surgery, it is necessary to continue observation, and considering the cost of the computed tomographic scanning it is necessary to determine the appropriate examination intervals.
As mathematical approach to this problem, based on the shape data of blood vessels, clarification of generating mechanism of aortic aneurysms has been performed. For this purpose, the estimation of probability by site within a vessel of aortic aneurysm is required. And to identify high-risk shape related to a diameter increase under the condition that occurs, need to create a discrimination model using them.
In this research, we smoothed and estimated the derivatives using local polynomials.
For bandwidth selection, we used a cross-validation method and Plug-in method. And, we compared each of the bandwidth selection and estimated of curvature and torsion.

CS05 ROBUST MODELLING AND HIGH-DIMENSIONAL DATA I

Session Chair: David Nott, National University of Singapore, Singapore
Venue: Seminar Room 3
Time: 17 Dec, 16:00-18:00

An asymptotic expansion of the distribution of the studentized linear discriminant function for large dimension

Takayuki Yamada
General Studies, College of Engineering, Nihon University.
E-mail: yma801228@gmail.com

Keywords: Linear discriminant rule; studentized statistic; (n; p) asymptotic; cut-off point.

Abstract: This paper is concerned with the problem of classifying a observation vector into one of two populations. Anderson (1973, Ann. Statist.) gave an asymptotic expansion of the studentized statistic, and derived cut-off point to achieve a specified probability of misclassification. But, as the dimension p becomes large, the precision of the approximation gets worse. In this paper, we proposed studentized statistic in terms of (n; p) asymptotic. An asymptotic expansion of the statistic is derived up to the order O_1, where $O_{j/2}$ is a term of j-th order with respect to $\{p^{-1/2}; N1^{-1/2}; N2^{-1/2}; m^{-1/2}\}$ for each sample

size N_i, m = N-p, N = N_1 + N_2-2. Using the expansion, we gave cut-off point to achieve a specified probability of misclassification.

Quantile coherence analysis

Yaeji Lim
Biostatistics and Clinical Epidemiology Center & Samsung Medical Centre.
E-mail: yaeji.lim@gmail.com

Keywords: Coherence analysis, Signal processing, Cross periodogram, Quantile regression.

Abstract: The coherence analysis measures the linear time-invariant relationship between two data sets and has been studied various fields such as signal processing, engineering, and medical science. However classical coherence analysis tends to be sensitive to outliers and focuses only on mean relationship. In this paper, we generalized cross periodogram to quantile cross periodogram and provide richer inter-relationship between two data sets. This is a general version of Laplace cross periodogram Li (2012) suggested. We prove its asymptotic distribution and compare them with ordinary coherence through numerical examples. We also present real data example to confirm the usefulness of quantile coherence analysis.

CS06 MARKOV CHAIN MONTE CARLO METHODS
Session Chair: Ajay Jasra, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 17 Dec, 16:00-18:00

Efficient strategy for the Markov chain Monte Carlo in high-dimension and its implementation

Kengo Kamatani
Osaka University and CREST, JST, Japan
Email: kamatani@sigmath.es.osaka-u.ac.jp.

Abstract: The purpose of this paper is to investigate efficient Markov chain Monte Carlo (MCMC) methods by simulation and high-dimensional asymptotic theory. Several MCMC methods are studied including the random-walk Metropolis (RWM), pCN (preconditioned Crank-Nicolson) and MpCN algorithms. These algorithms have reversible proposal transition kernels. Key fact is that the performance is heavily depends on the relation between the target distribution and the invariant distribution of the proposal transition kernel. We illustrate that MpCN is the best in terms of the convergence rate. Some practical implementation issue is discussed.

Ice core dating using the particle Markov chain Monte Carlo method

S. Nakano,
The Institute of Statistical Mathematics, School of Multidisciplinary Science, SOKENDAI,

Abstract: Ice cores provide fundamental information on the climatic changes over the past hundreds of thousands of years. In order to make use of the information from ice cores, it is important to accurately estimate the relationship between age and depth in the ice cores. The age-depth relationship depends on the accumulation of snow at the site of the ice core and the thinning process due to the horizontal stretching and vertical compression of an ice layer. In estimating the age as a function of depth, it is necessary to simultaneously estimate the parameters describing the accumulation and thinning processes which cannot be represented by linear equations. We propose a new dating technique which estimates both the age-depth relationship and the related parameters. The estimation is achieved using the particle Markov chain Monte Carlo method, which is a hybrid method combining a sequential Monte Carlo method and the Markov chain Monte Carlo method. The use of the particle Markov chain Monte Carlo method enables us to estimate the parameters without assuming linearity. It is demonstrated how this dating technique works by applying it to the ice core data from Dome Fuji in Antarctica.

MCMC simulation methods for high-dimensional Gaussian graphical models using precision matrix

Niharika Gauraha
Systems Science and Informatics Unit (SSIU), Indian Statistical Institute (ISI) Bangalore
E-mail: niharika.gauraha@gmail.com

Keywords: Gaussian Graphical Models, MCMC, Data Simulation, Metropolos-Hastings, MALA

Abstract: We propose new applications of Metropolis-Hastings (M-H) MCMC simulation methods to generate a representative data sample for a high dimensional Gaussian Graphical Model (GGM) given its graph structure. When the data set follows a multivariate normal (MVN) distribution, i.e. $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p}) \sim \mathrm{N}_\mathrm{p}(\mu; \Sigma)$. The precision matrix $\mathrm{P} = \Sigma^{-1}$ can be directly translated into a GGM, i.e. $\mathrm{P}_{\mathrm{ij}} = 0, \mathrm{i} \neq \mathrm{j}$ implies that X_i and X_j are independent given the rest. Similarly given the graph structure of a GGM, we can derive a representative precision matrix.
The unnormalized MVN density is expressed in terms of P as

$$\phi(\mathrm{x}) \propto \exp\{-\frac{1}{2}(\mathrm{x}-\mu)'\mathrm{P}(\mathrm{x}-\mu)\}$$

Using the above density as the target function we propose the following ways to simulate from MVN.
1. The random walk M-H based method using the efficient symmetric jumping kernel (scaled i.i.d. univariate Gaussian).
2. The Metropolis Adjusted Langevin algorithm (MALA) based simulation.
The experimental results show that our algorithm produces representative data sets for different sizes and variety of graph topology.

Population-based MCMC method for dynamic generalized linear models

Guangbao Guo
School of Mathematics, Shandong University, Jinan 250100, China

Abstract: In this paper, we consider one population-based MCMC (Pop-MCMC) method for dynamic generalized linear models (DGLMs), Pop-MCMC can be described as generating a collection of random variables in parallel in order to simulate from some target density. The method is important as many challenging sampling problems which cannot be dealt with successfully by general MCMC methods. Furthermore, we give some several optimal properties of our proposed method, under some technical conditions. At last, some simulation results are reported to illustrate the proposed method, we provide a comparison of the methods (KS, Gibbs, MH and Pop-MCMC) about the fitting error criteria, when the models are with Poisson-binomial distribution. Experimental results have confirmed that Pop-MCMC learn more efficiently than those methods with no information exchange, and the MCMC technique often produces satisfactory performance.

CS07 ROBUST MODELLING AND HIGH-DIMENSIONAL DATA II
Session Chair: Alex Beskos, University College of London, UK
Venue: Seminar Room 3
Time: 18 Dec, 10:20-12:20

The improved Value-at-Risk for heteroscedastic processes and their coverage probability

Khreshna Syuhada
Statistics Research Division, Institut Teknologi Bandung, Indonesia
Email: khreshna@math.itb.ac.id

Keywords: Autoregressive, conditional heteroscedasticity, coverage probability, estimator variability, time series forecasting

Abstract: A risk measure commonly used in financial risk management, namely Value-at-Risk (VaR), is studied. In particular, we find a VaR forecast for heteroscedastic processes such that its (conditional) coverage probability close to the nominal. To do so, we pay attention to the effect of estimator

variability such as asymptotic bias and mean square error. Numerical analysis is carried out to illustrate this calculation for Autoregressive Conditional Heteroscedastic (ARCH) model, an observable volatility type model. In comparison, we do finding VaR for latent volatility model i.e. Stochastic Volatility Autoregressive (SVAR) model. It is found that the effect of estimator variability is significant to obtain VaR forecast with better coverage. In addition, we may only be able to assess unconditional coverage probability for VaR forecast of SVAR model. This is due to the fact that the volatility process of the model is unobservable.

Ridge regression based on MML estimators with LTS error distributions

Sukru Acitas* and Birdal Senoglu[#]
*Anadolu University, Department of Statistics, 26470 Eskisehir, Turkey
E-mail: sacitas@anadolu.edu.tr.
[#]Ankara University, Ankara, Turkey

Keywords: Ridge regression, modified maximum likelihood, mean square error, Monte-Carlo simulation.

Abstract: In the linear regression model, some assumptions are expected to be held to make statistical inference. It is well known that violence of one of the assumptions causes serious problems in the estimation and the hypothesis testing procedures. In this study, we deal with simultaneous violence of two assumptions in detail: Multicollinearity in the exploratory variables and non-normality of the errors terms. Therefore, we develop a methodology to overcome both multicollinearity and non-normality problems. Since ridge regression (Hoerl & Kennard, 1970 and Hoerl et al., 1975) is widely used method for solving multicollinearity problem, we here provide new ridge estimators by using Tiku's (1967, 1968) modified maximum likelihood (MML) methodology under the assumption of long-tailed symmetric (LTS) error distributions. A comprehensive Monte-Simulation study is conducted for comparing the efficiencies of proposed estimators and ordinary ridge estimators which are based on least squares (LS), see i.e. M Donald & Galarneau (1975) and Kibira (2003). The results show that ridge estimators based on MML estimators give satisfactory results in terms of having smaller mean square error (MSE) values. It should also be noted that as far as we know this is the first study considering ridge regression with MML methodology.

An approach to the modeling of abstraction systems for collective text

Ken NITTONO
Hosei University

Abstract: With the advance in quality of internet as infrastructure and its service as application, the amount of various kinds of data, which are

exchanged through the services, such as commercial transaction log or text messages on social networks has been increasing its volume.
Especially, as the result of the spread of ubiquitous use of mobile phones and tablet devices, the condition of the increase has been enhanced.
In those cases of processing such an enormous amount of data, the effective utilization of statistical analyzing approaches such as data mining methods becomes essential to obtain significant results in relatively short range of time.
In this case, we deal with text data particularly and treat abstraction systems that extract useful information from such a large amount of data.
A modeling approach for constructing systems with abilities of searching and selecting meaningful contents throughout the original data and also accumulating the selected contents efficiently or systematically for the further succeeding use of them is studied.
In addition, the necessity and the importance of user interface for the use of the systems from mobile phones or tablet devices are also mentioned.

Statistical Properties of Random Sampling Regression Based on Mahalanobis Distance for Big Data

Zhen Yan
School of Statistics, Renmin University of China.
E-mail: zhenyan@ruc.edu.cn

Keywords: Big Data, Random Sampling, Mahalanobis Distance, Statistical Distance, Weighted Least Square Estimation, Computational Efficiency.

Abstract: In the era of Big Data, facing with massive amounts of data, valid statistical analysis method for Big Data is becoming increasingly important. One of popular methods in dealing with big data is sampling. In this approach, one can utilize only a small portion of the data to estimate the parameter of interesting. Ma et al. (2013) firstly focused on the statistical quality of the leverage-based sampling algorithm. Motivated by the thought of leveraging sampling algorithm, we propose some sampling algorithms based on Mahalanobis distance in this paper. We also derive the statistical properties, including bias and variance, both conditional and unconditional on the observed data. Furthermore, we extend the sampling method based on Mahalanobis distance to the more efficient sampling method based on statistical distance. The simulation results show that our methods are more efficient and the real data analysis confirms the performance of the estimators.

CS08 STATISTICAL METHODOLOGY I
Session Chair: Gan Fah Fatt, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 18 Dec, 10:20-12:20

Looking beyond the financial ratios through data envelopment analysis

Amritpal Singh Dhillon
Hemchandracharya North Gujarat University, Gujarat, India
E-mail: amritpal_dhillon2000@yahoo.com

Keywords: banking industry, output benchmarking, target evaluation, scale inefficiency.

Abstract: In normal phenomenon banks performance is measured with the help of financial ratios like return on assets, liquidity ratios or return on equity etc. giving extreme importance to monetary items. On other hand, it's quite difficult to measure non-monetary items over a period of time. Apart from this problem, one is interested in knowing best performed year for an organization. This paper has been formulated and targeted to overcome such barriers, so that complete comprehensive picture of the organization performance can be revealed and proper policies can be formulated to achieve desired goals. Data Envelopment Analysis (DEA), a non-parametric approach had been used very extensively in various fields over the past two decades. Same mathematically tool is utilized in evaluating the performance of India's leading bank i.e. State Bank of India over a period of 10 years. The objective of this paper is to identify the relatively best performing year based on monetary as well as non-monetary values. Secondary objective is to highlight the impact of various inputs over the output. Under different formulated Models A, B and C the output can be increased by 0.5%, 8.2% and 5.4% respectively with the same inputs. Scale efficiency results shows that financial year 2005-06 and 2006-07 was the golden time for the bank to expand their banking activities in order to have scale leverage.

Proportional odds of nutritional status of under-five children in Nigeria indexed by MUAC

Anthony Ekpo
Dept. of Mathematics/Statistics/Computer Science
University of Agriculture, Makurdi, Nigeria
Email: anthony.ekpo@uam.edu.ng

Keywords: Statistical surveillance, malnutrition, severe acute malnutrition, mid-upper arm circumference, anthropometrics, emergency nutrition assessment and bilateral edema.

Abstract: This paper presents results regarding the impacts of some anthropometric, epidemiological and demographic factors on the nutritional status of under-five children which were categorized into three ordinal groups of Severe Acute Malnutrition (SAM), Moderate Acute Malnutrition (MAM) and Global Acute Malnutrition (GAM) in Kazaure Local Government Area of Nigeria. An ordinal logit model that depicted the log-odds in favor of GAM (normal) child was fitted to the data based on a surveillance indexed by Mid-Upper Arm Circumference (MUAC). By this, the proportional odds of being in either of the nutritional status based on age, sex and measles when malnutrition surveillance is indexed by MUAC were determined. The results showed that, the proportional odd of measuring malnutrition prevalence using the MUAC index is (OR = 1.138, with $p < 0.001$). The sex of a child does not play a major role in determining the nutritional status of a child. Being vaccinated for Measles did not play a major role in classifying a child's malnutrition status. Rather, variables that has to do with access to potable water, access to household food for children and other socioeconomic variable could be considered. Edema as a morbid state was found to be redundant in the study.

Anoa: extreme price fluctuation monitoring and reporting application with Tukey algorithm

Brilian Surya Budi
Department of Computational Statistics, Sekolah Tinggi Ilmu Statistik, East Jakarta, DKI Jakarta, Indonesia

Keywords: Scheduled data crawler, Tukey Algorithm, fluctuation price, real time monitoring and reporting

Abstract: Many countries have a fluctuation on price of every commodities. This fluctuation cloud be on significant data or outlier. It is a big question, how to determination whether it is on tolerable condition or it is should be handed as special condition at the real time. Anoa is an alternative to answer the question. Anoa is the web-based application can monitor the price fluctuation at real time and it can give any information if there is anomalous price increase or decrease. First, it using a scheduled data crawler for retrieve data from website daily which provides data of commodities with extreme price fluctuation. Then to determine if the increase or decrease of price is still in the tolerably level, Tukey Algorithm is applied. In Indonesia, those commodities are rice grains, red chilies, sugar, cooking oil, and meats. Extreme price increase and decrease frequently occurs on those commodities at certain periods. Furthermore, this application can be considered as a tool for defining a fluctuation price by Indonesian Government to keep the price stable.

On the prediction of election results at early stages of counting

RAB Abeygunawardana,

Department of Statistics, University of Colombo, Sri Lanka
Email: rab_abey@stat.cmb.ac.lk

Keywords: inequality constraints, least squares estimation, election results, auxiliary information

Abstract: Predicting election results accurately is a big challenge for a country as well as a good opportunity for a statistician. Current methods used to predict election results are mainly based on prior knowledge about the election. Attention is not paid to predict by combining previous results and partially released results on the election night. Here we propose to use inequality constrained least squares estimation to predict the election results by combining results released at early stages of counting and auxiliary information obtained from past elections and the current election. Constraints are determined using past election results and expert opinion. Based on correlations with the total votes (Y) received by a leading party with number of votes received by that party in the immediate previous election (x1) and number of registered voters in current election (x2) were selected as auxiliary variables. These variables were first transformed to overcome the problem of multicollinearity. Predicted results are then compared with actual results using appropriate statistical measures.

CS09 SEMIPARAMETRIC METHODS
Session Chair: Scott Sisson, University of New South Wales, Australia
Venue: Seminar Room 2
Time: 18 Dec, 13:30-15:30

A modification of the Liu regression estimators in partially linear models

Gulin Tabakan
Department of Mathematics, Aksaray University, 68100, Aksaray, Turkey
gtabakan@aksaray.edu.tr

Keywords: Difference-based estimator; Liu estimator; Ridge estimator; Multicollinearity

Abstract: In this paper, we introduce a new difference-based regression estimator called difference-based modified Liu estimator based on prior information for the vector of parameters in a partially linear model (PLM) as an alternative to the classical difference-based regression estimator in the existence of multicollinearity problem. The theoretical properties of the proposed estimator and its relationship with some existing difference-based biased methods designed for PLM are investigated. Finally, a Monte Carlo simulation is done to illustrate some of the theoretical results.

Selection of variable and classification boundary for functional data by logistic regression

Hidetoshi Matsui
Faculty of Mathematics, Kyushu University.
E-mail: hmatsui@math.kyushu-u.ac.jp

Keywords**:** Functional data analysis, Lasso, Model selection, Regularization

Abstract: Penalties with a ℓ1 norm provide solutions in which some coefficients are exactly zero and can be used for selecting variables in regression settings. When applied to the logistic regression model, they also can be used to select variables which affect classification. We focus on the form of ℓ1 penalties in the logistic regression models for functional data, in particular, their use in classifying functions into three or more groups while simultaneously selecting variables or classification boundaries. By extending the ℓ1=ℓq penalties, we propose a new class of penalties in order to appropriately estimate and to select variables or boundaries for the functional multiclass logistic regression model. The parameters involved in the model is estimated by the framework of the blockwise descent algorithm, and then a value of the tuning parameter included in the regularization method is decided by a model selection criterion. Analysis of real data show that the form of the penalty should be selected in accordance with the purpose of the analysis.

Initial value selection of the EM algorithm for Gaussian mixture models

Masahiro Kuroda
Department of Socio-Information, Okayama University of Science.
E-mail: kuroda@soci.ous.ac.jp

Keywords: EM algorithm; vector ϵ algorithm; Acceleration of convergence; Re-starting; Initial value selection

Abstract: The EM algorithm is a standard tool for maximum likelihood estimation in Gaussian mixture models. Then the choice of initial values can heavily influence the speed of convergence of the EM algorithm. The solution of the EM algorithm can also higher depend on these values. Many initialization methods have been proposed, but there is not a best strategy that outperforms the rivaling procedures in all cases. The standard procedure for solving the initial value selection problem is the multi-starting approach by generating random numbers. Biernacki et al. (2003) provided the random initialization method using short runs of the EM algorithm (em-EM). We use the em-EM algorithm as the initial value selection method and apply the vector ϵ algorithm to speed up the convergence of the em-EM algorithm. Then we propose a stopping condition using the Aitken _2 method for reducing the number of iterations and computational time of the vector ϵ acceleration of the em-EM algorithm. When obtaining the best initial value, the vector ϵ acceleration can be re-applied to the EM estimation of the parameters of Gaussian mixture models. Furthermore, we improve the speed of

convergence of the vector ϵ acceleration of the EM algorithm using a re-starting procedure given in Kuroda et al. (2015).

Solving fused group lasso problems via block splitting algorithms

Tso-Jung Yen
Institute of Statistical Science, Academia Sinica, Taiwan

Keywords: Fused lasso; Group lasso; Scalability; Alternating direction method of multipliers; Block splitting algorithms.

Abstract: In this paper we propose a distributed optimization-based method for solving the fused group lasso problem, in which the penalty function is a sum of Euclidean distances between pairs of parameter vectors. As a result of that, the penalty function is not separable in terms of these parameter vectors. To make the penalty function separable, one common way is to introduce a set of auxiliary variables that represent the differences between pairs of parameter vectors. This representation can be seen as a linear operator on the joint vector of the parameter vectors, and the resulting augmented Lagrangian will have a coupling quadratic term involving the linear representation. Even though the linear representation is separable in terms of the parameter vectors, the coupling quadratic term is not. To make the coupling quadratic term separable, we further introduce a set of equality constraints that connect each parameter vector to a group of paired auxiliary variables. With these newly introduced equality constraints, we are able to derive a modified augmented Lagrangian that is separable either in terms of the parameter vectors or in terms of the paired auxiliary variables. This separable property further facilitates us to solve the fused group lasso problem by developing an iterative algorithm with that most tasks can be carried out independently in parallel. We evaluate performance of the parallel algorithm by carrying out fused group lasso estimation for regression models using simulated data sets. Our results show that the parallel algorithm has a massive advantage over its non-parallel counterpart in terms of computational time and memory usage. In addition, with additional steps in each iteration, the parallel algorithm can obtain parameter values almost identical to those obtained by the non-parallel algorithm.

CS10 COMPLEX MULTIVARIATE MODELS II
Session Chair: Hongmei Zhang, University of Memphis, USA
Venue: Seminar Room 3
Time: 18 Dec, 13:30-15:30

The effect of responses missing at random on the optimal cohort design of linear mixed effect models

Kim May Lee, Stefanie Biedermann, Robin Mitra

University of Southampton (UK)

Keywords: Optimal design, missing data, linear mixed effects models, pairwise deletion

Abstract: In many areas, linear mixed effects models are employed in the analysis of longitudinal data where each experimental unit has repeated measurements on the same outcome collected at several points in time. It is often unavoidable to have missing observations in the data set especially when the duration of the study is long. Most of the literature on experimental design focus on finding the optimal allocation of time points, with the assumption of having fully observed data. Moreover, the missing data analysis method is often ignored by the experimenters at the design stage of an experiment. Following Ortega-Azurduy, Tan and Berger (2007), we extend the optimal design framework for responses missing at random, to account for having more than one cohort of subjects in the longitudinal study. Pairwise deletion is assumed to be implemented in the estimation of the fixed effect parameters. An extra covariate is incorporated into the framework to provide a more flexible set-up to the applications. An illustration of D-optimal cohort designs will be presented for a polynomial mixed effects model.

Bayesian method for testing differential directed acyclic graphs

Hongmei Zhang,
Division of Epidemiology, Biostatistics, and Environmental Health Sciences, School of Public Health, University of Memphis, Memphis, TN 38017. USA.

Abstract: Graphical models are essential to describe networks among different factors. Various methods to construct graphs have been proposed. In practice, it is critical to assess the agreement between networks constructed under different treats, e.g., a network formed by a large number of epigenetic factors (e.g., DNA methylation) among subjects who smoke versus that among non-smokers. There is rather limited effort in this area. We propose a Bayesian method to build directed acyclic graphs (DAGs) based on a given order of nodes and simultaneously test the agreement between DAGs. The network construction and differential graphs testing are built upon the concept of variable selection. Simulations demonstrate the applicability of the method and a real data application to an epigenetic data is implemented to illustrate the method.

Vector regression without specifying marginal distributions or association structures

Alan Huang
University of Queensland, Australia

Abstract: We introduce a flexible yet parsimonious framework for vector regression based on nonparametric multivariate exponential families. The key

feature is that underlying exponential family can be left completely unspecified in the model and can be estimated nonparametrically from data along with the usual regression coefficients using a maximum empirical likelihood approach. Its usefulness in practice is demonstrated via various simulations and data analysis examples.

CS11 EXPLORATORY AND GRAPHICAL METHODS
Session Chair: Alex Thiery, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 18 Dec, 13:30-15:30

Constrained asymmetric MDS based on raidus model

Kensuke Tanioka
Graduate School of Culture and Information Science, Doshisha University.
E-mail: eim1001@mail4.doshisha.ac.jp

Keywords: Skew-symmetric data; Clustering; Dissimilarity

Abstract: Asymmetric dissimilarity data is defined as dissimilarity describing the asymmetric relations between objects, and is observed in various fileds such as marketing research. The dissimilarity from object i to object j is not necessarily to the same as that from object j to i. For example in marketing research, it is important to interpret competitive relations among brands from asymmetric dissimilarity data for the competitive relations of brands. To describe the relations, asymmetric MDS (Borg and Groenen, 2005; Chino, 2012; Saito and Yadohisa, 2005) is one of visualizing methods. However, improving technology provides us to deal with large asymmetric dissimilarity data. In the situation, it becomes difficult to interpret the relations between objects for the asymmetries. To overcome the problem, we propose a new asymmetric MDS based on the radius model (Okada and Imaizumi, 1987), given the information for classes of objects. There are two features of the proposed method. First, the asymmetric relations are described with the small number of parameters. Second, relations between classes are interpreted easily because objects belonging to the same class are located closely by the constraints of the proposed method.

Bayesian asymmetric multidimensional scaling for two-mode three-way count data by using log-linear model

Jun Tsuchida
Graduate School of Culture and Information Sciences, Doshisha University.

Keywords: Asymmetric MDS, Bayesian imputation

Abstract: Given (dis)similarity data between objects, multidimensional scaling (MDS) is one of the methods for describing the relations between objects. When (dis)similarity data is asymmetry, asymmetric MDS represents the relations based on asymmetry. In many cases, asymmetric (dis)similarity data is often obtained as count data which records the number of occurrences of a particular event. For example, brand switching data is the data which represents the number of switching from one brand to another. De Rooij and Heiser (2003) have proposed asymmetric MDS for count data. When we obtain asymmetric (dis)similarity data in some sources, which is called Two-mode Three-way data, we cannot apply the asymmetric MDS proposed by De Rooij and Heiser to this data. Moreover, Two-mode Three-way data has missing values in many cases. To overcome these problems, we propose the Bayesian asymmetric MDS by using log-linear model. By applying Bayesian estimation, we can impute missing data and utilize prior knowledge.

Neural network control chart for monitoring the individual measurements

S. M. Nimbale
Department of Statistics, Solapur University, Solapur (MS), India.
Email: smnimbale@sus.ac.in

Keywords: Control chart, individual control chart, average run length, artificial neural network.

Abstract: In quality control practice, situations frequently arise that require a charting procedure for individual measurements. Charting of individual observations has received extensive attention in recent day. The statistical performance of individual control chart is depends on the normality assumption. Nevertheless, there has always been some concern about this normality assumption. Especially when individual measurements are used, the normality assumption is risky. In recent literatures ANN models were proposed as alternative tool for various types of control charts as data independency and normality is not an assumption in ANN models. In this proposed work we compared the performance of Shewhart X individual chart and Shewhart Moving Range charts with the performance of Neural Network Control Chart in term average run length (ARL) under normal and non-normal distributions. This simulated study indicates neural network control chart is efficient to detect shifts in process of single observation data under normal and non-normal distribution.

CS12 STATISTICAL METHODOLOGY II
Session Chair: Zhou Yan, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 18 Dec, 15:50-17:50

Study on comparison between run survey datasets of air dose rate

Hiroki Takamaru
Graduate School of Information Science and Technology, Hokkaido University, JAPAN

Keyword: Big data, Fukushima prefecture

Abstract: We analyze the two different run survey datasets in Fukushima prefecture, Japan and discuss the differences and their interpretation.
Air dose rates are measured by various methods including run survey after Fukushima Daiichi nuclear accident. Run survey measures gamma ray on road by vehicle. In this study, we use the two run survey datasets: One is the dataset measured by cars and recorded an air dose rateper100m*100 m mesh. The other is measured by buses and much recorded per mesh. We can make comparative discussion of both datasets since their meshes correspond. Through the comparison and the analysis, we get to detect several patterns on the difference between the two and they might have a relation to the land usage.

Semiparametric inference under nonignorable nonresponse

Kosuke Morikawa
Graduate School of Engineering Science, Osaka University, Japan.
E-mail: morikawa@sigmath.es.osaka-u.ac.jp

Keywords: Incomplete Data, Nonresponse, Nonignorable Missingness.

Abstract: Nonresponse has become a major problem in many fields of empirical studies. When a missing-data mechanism is not at random (NMAR) or nonignorable, estimators even for mean or median such as a sample mean or a sample median may be severely biased. Moreover, to analyze such data by conventional methods, they often need a distributional assumption on a conditional distribution of a response variable given covariates as well as that on a missing-data mechanism. Specifying the conditional distribution is very difficult since one must do it only from the information of observed part of the sample, and if it was misspecified, estimators obtained under the wrong assumption would be biased although specifying it is not our inter- est. Therefore, we propose kernel-based semiparametric estimators for the parameter without postulating any distributional assumptions on it. Some asymptotic properties of our estimators are derived. Results of a simulation study and real data analysis are also presented.

The median estimator of quantile regression

Ryunosuke Tanabe
Graduate School of Engineering Science, Osaka University
E-mail: tanabe@sigmath.es.osaka-u.ac.jp

Keywords: spike and slab, quantile regression, model selection.

Abstract: The quantile regression is used to estimate the conditional distribution of a response variable and one of the advantage is in the model robustness because we do not need to assume any distribution. Like the Bayesian penalty mean regression, a penalty quantile regression estimator and its loss funciton can be interpreted as a posterior mode estimator and an asymmetric Laplace distribution, respectively. Also an L1 penalty quantile regression estimator can be interpreted as a posterior mode if coefficient parameters are independent of both double exponential prior and an asymmetric Laplace likelihood. The Bayesian penalty quantile regression provides an interval estimator, but its coefficient estimator is not always zero. On the other hand, The use of a spike and slab prior can make the posterior median estimator zero exactly, so that we have better method and estimators than the Bayesian penalty quantile regression. The posterior median estimator also has the consistency of variable selection if the design matrix is orthogonal.

Sample size calculation for model selection

Shinpei Imori
Graduate School of Engineering Science, Osaka University
E-mail: imori@sigmath.es.osaka-u.ac.jp

Keywords: Model selection; Sample size calculation

Abstract: Model selection by information criterion (IC) is an important topic in real data analysis. For example, in the simulation for evaluating cancer risk, by specializing the essential risk factors gives us optimal model, which describes the mechanism of carcinogenesis. Here we have to note that the result by IC is a random variable, because IC is calculated from observed data.
There are some asymptotic studies of the selected model by IC. However, the sample is finite in practical, and to collect data for satisfying the sufficient number is not easy. Then, it is important to evaluate the uncertainly of result from finite sample since the result for a different dataset may differ. Thereby, the approximation of the selection probability in finite sample plays a key role. We propose an approximation procedure for the selection probability of the true model. The selection algorithm in our idea is based on Imori et al. (2014, TR 14-01, Statistical Research Group, Hiroshima University). This algorithm leads a simple formulation of the approximation and relaxes assumptions for unknown parameters in the true model. Our procedure gives how to design a sample size for adequate model selection.

CS13 STATISTICAL MODELLING I
Venue: Seminar Room 2
Time: 19 Dec, 9:00-11:00

Mathematical models for radiotherapy based on dose volume histogram

Masahiro Mizuta
Advanced Data Science Laboratory, Information Initiative Center, Hokkaido University.
E-mail: mizuta@iic.hokudai.ac.jp

Keywords: Cell Survival Model, Optimization, Tumor

Abstract: We focus on the optimization of planning in radiotherapy. Radiotherapy plays an important role in the treatment of solid tumors. The principle of Radiotherapy is to kill cancer cells and minimize damage effect on OAR (organs at risk). We express this principle mathematically using cell survival models.
Mizuta et al. (2012) proposed a mathematical method for selecting a single or fractionated irradiation regime based on physical dose distribution. But, the method does not consider dose volume histogram (DVH), which represents the distribution of dose on the OAR or Tumor. We extend LQ model to volume model and evaluate a planning of radiotherapy.

Spatial Bayesian hierarchical model with variable selection to fMRI data

Kuo-Jung Lee
Department of Statistics, National Cheng-Kung University.
E-mail: kuojunglee@mail.ncku.edu.tw

Keywords: Bayesian, fMRI, SGLMM

Abstract: We propose a spatial Bayesian hierarchical model to analyze functional magnetic resonance imaging data with complex spatial and temporal structures. Several studies have found that the spatial dependence not only appears in signal changes but also in temporal correlations among voxels. However, currently existing statistical approaches ignore the spatial dependence of temporal correlations for the computational efficiency. We consider the spatial random effect models to simultaneously model spatial dependences in both signal changes and temporal correlations, but keep computationally feasible. Through simulation, the proposed approach improves the accuracy of detection of brain activities. We study the properties of the model through its performance on simulations and a real event-related fMRI data set.

Optimizing Two-level Orthogonal Arrays for Estimating Main Effects and Pre-specified Two-factor Interactions

Ping-Yang Chen
Department of Statistics, National Cheng Kung University.
E-mail: R28021017@mail.ncku.edu.tw

Keywords: D-optimal designs; Hadamard matrices; orthogonal arrays; particle swarm optimization.

Abstract: In this work, we are interested in constructing D-optimal orthogonal arrays that allow joint estimation of main effects and some specified two-factor interactions. A hybrid algorithm that combines an enhanced stochastic evolutionary algorithm and a discrete particle swarm algorithm is proposed. We demonstrate the performance of the proposed algorithm by generating orthogonal arrays with different interaction set-ups, and then comparing with these designs obtained in the previous literatures. In addition to the numerical generating algorithm, we also study the upper bound of the determinant of the corresponding matrix for D-optimality based on our numerical results. Joint with Ray-Bing Chen (NCKU) and Chunfang Devon Lin (Queen's University)

Estimation of scale-free networks with the exponentiation of minimax concave penalty

Kei Hirose,
Division of Mathematical Science, Graduate School of Engineering Science, Osaka University.
E-mail: hirose@sigmath.es.osaka-u.ac.jp

Keywords: Gaussian graphical model, lasso, scale-free networks, mini-max concave penalty

Abstract: We consider the problem of sparse estimation of undirected graphical models via the L1 regularization. The ordinary lasso encourages the sparsity on all edges equally likely, so that all nodes tend to have small degrees. On the other hand, many real-world networks are often scale-free, where some nodes have a large number of edges. In such cases, a penalty that induces structured sparsity, such as log penalty, performs better than the ordinary lasso. In practical situations, however, it is difficult to determine an optimal penalty among the ordinary lasso, log penalty, or somewhere in between. In this paper, we introduce a new class of penalty that is based on the exponentiation of the minimax concave penalty. The proposed penalty includes both the lasso and the log penalty, and the gap between these two penalties is bridged by a tuning parameter. We apply the cross-validation to select an appropriate value of the tuning parameter. Monte Carlo simulations are conducted to investigate the performance of our proposed procedure. The numerical result shows that the proposed method can perform better than the existing log penalty and the ordinary lasso.

Poster Session

Modelling liquidity supply in limit order book with a vector functional autoregressive (VFAR) model

Wee Song Chua

Department of Statistics & Applied Probability, National University of Singapore E-mail: a0054070@u.nus.edu

Keywords: Bid and Ask Supply Curves; High Frequency Financial Time Series; Sieve Estimator

Abstract: We propose a Vector Functional Autoregressive (VFAR) model to describe the dynamics of the liquidity supply in the limit order book. Given the time series of two curves, we conduct expansion based on the B-splines and develop a consistent Sieve estimator within the VFAR framework. Applying the proposed model to the bid and ask supply curves of ten stocks traded at the National Association of Securities Dealers Automated Quotations (NASDAQ) stock market in 2015, we show that the VFAR model provides good accuracy and interpretability.

Predict electricity price curves with Fisher-Rao distance

Jiejie Zhang
Department of Statistics & Applied Probability, National University of Singapore.
E-mail: jiejiezhang@u.nus.edu

Keywords: Functional time series; warping function; Karcher mean

Abstract: We propose a new model to predict electricity price curves that accounts for the time phase and level amplitude variations of the functional time series simultaneously. In particular, we adopt the Fisher-Rao distance to convert the data into warped functions and time-dependence warping functions. PCA is implemented to extract essential factors for the warped functions, while FAR helps to explain the dynamic structure of the warping functions. We implement the predictive model to the California market.

Quantile variable selection for high dimensional data analysis

Muhammad Amin
School of Mathematical Sciences, Dalian University of Technology, Dalian, 116023, P.R.China
Email: aminkanju@gmail.com
Keywords: Quantile Regression, SCAD, Variable Selection, High Dimensional Data

Abstract: The present quantile variable selection methods are only applicable to limited number of predictors or do not have oracle property associated with estimator. It is considered as another method to ordinary least squares (OLS) regression in case of the outliers and the heavy tailed errors existing in linear models. The variable selection through quantile regression with diverging number of parameters is studied in this work. The convergence rate of this estimator with smoothly clipped absolute deviation (SCAD) penalty function is also discussed. Moreover, the oracle property with proper selection of tuning parameter for quantile regression under certain regularity conditions is also established. Moreover, the rank correlation screening method (RCS) is used to accommodate ultra-high dimensional data settings. Monte Carlo simulations demonstrate finite performance of the proposed estimator. The results of real data reveal that this approach provides substantially more information as compared to OLS, conventional quantile regression (RQ) and quantile lasso (Q-Lasso).

Sparse principal component regression modeling and generalized information criterion

Heewon Parka
Faculty of Global and Science Studies, Yamaguchi University, Japan.
E-mail: hwpark@yamaguchi-u.ac.jp

Keywords: Adaptive regularization; Information criterion; Principal component regression; sparse regression modeling

Abstract: The components selection is a vital matter in principal components analysis. Relatively little attention, however, was paid to this issue, and the existing studies for principal component analysis were based on ad-hoc methods. We propose a novel strategy for principal component selection via sparse principal component regression modeling. In order to effectively perform for principal component selection, we consider adaptive L1-type penalty based on singular values of components, and propose adaptive penalized principal component regression. The proposed method can incorporate explanation power of components to principal component selection procedure. In sparse regression modeling, choosing the regularization parameters is a crucial issue, since variable selection and model estimation heavily depend on the selected regularization parameters. We derive a model selection criterion for choosing the regularization parameters of the proposed adaptive L1-type regularization method in line with the generalized information criterion. Monte Carlo simulations indicate that the proposed strategies outperform for principal component regression modeling.

Looking beyond the financial ratios through data envelopment analysis

Amritpal Singh Dhillon
North Gujarat University, Gujarat, India
E-mail: amritpal_dhillon2000@yahoo.com

Keywords: banking industry, output benchmarking, target evaluation, scale inefficiency.

Abstract: In normal phenomenon banks performance is measured with the help of financial ratios like return on assets, liquidity ratios or return on equity etc. giving extreme importance to monetary items. On other hand, it's quite difficult to measure non-monetary items over a period of time. Apart from this problem, one is interested in knowing best performed year for an organization. This paper has been formulated and targeted to overcome such barriers, so that complete comprehensive picture of the organization performance can be revealed and proper policies can be formulated to achieve desired goals. Data Envelopment Analysis (DEA), a non-parametric approach had been used very extensively in various fields over the past two decades. Same mathematically tool is utilized in evaluating the performance of India's leading bank i.e. State Bank of India over a period of 10 years. The objective of this paper is to identify the relatively best performing year based on monetary as well as non-monetary values. Secondary objective is to highlight the impact of various inputs over the output. Under different formulated Models A, B and C the output can be increased by 0.5%, 8.2% and 5.4% respectively with the same inputs. Scale efficiency results shows that financial year 2005-06 and 2006-07 was the golden time for the bank to expand their banking activities in order to have scale leverage.

Holonomic properties for the distribution of the sample correlation coefficient

Hiroki Hashiguchi
Tokyo University of Science
E-mail: hiro@rs.tus.ac.jp

Keywords: Groebner bases, Normal population, Partial differential equation

Abstract: We focus on the holonomic properties of the pdf (probability density function) and cdf (cumulative distribution function) of the sample correlation coefficient under normal population. At first, we present that the pdf and cdf of sample correlation coefficient are holonomic functions. Some calculation on the annihilate ideal for the pdf and cdf can be conducted by HolonomicFunctions.m package on Mathematica. Secondly, we obtain the exact initial values on the holonomic gradient method for the pdf. Finally, we show the inverse function of the cdf has a nonlinear differential equation, therefore we can utilize numerical methods to calculate it.

Approximate gamma distributions for eigenvalues of a complex Wishart matrix and applications of MIMO capacity

Tatsuya Kuwabara and Hiroki Hashiguchi
Tokyo University of Science
E-mail: 1415603@ed.tus.ac.jp

Keyword: complex nonsingular Wishart distribution, complex singular Wishart distribution, MIMO system, approximate gamma distribution, approximate product of gamma distribution

Abstract: We discuss the asymptotic distribution for the eigenvalues of complex singular and nonsingular Wishart matrices, and their applications to MIMO system. Our results are similar to the real Wishart matrix each of whose eigenvalue asymptotically distributes as a chi-square distribution. We propose two kinds of approximate gamma distributions for the complex case. The first approximation is derived from using a gamma distribution with suitable parameters. The second approximation is based on the product of gamma distributions. We compare them with empirical distribution by Monte Carlo simulation. We also calculate MIMO capacity using the proposal distributions.

Consumer consciousness on domestic and imported food

Yumi Asahi
Department of Management Systems Engineering, Tokai University
E-mail: asahi@tsc.u-tokai.ac.jp

Keywords: questionnaire surveys, domestic food, the factorial analysis, structural equation modeling analysis

Poster

Abstract: In Japan, the production of food has fallen from the latter half of the 1980's because of decreasing of agriculture workers. On the other hand, the amount of imported food has increased, because the price of imported food is cheap and the stable supply of them is possible. Our country, needs of domestic food have risen because consumers have become increasingly aware of problem related to reliable standard of food. This paper analyzed purchasing behaviors of domestic food and imported food through a face to face interviews survey of housewives in Japan. This paper analyzed purchasing behaviors of domestic food and imported food through two questionnaire surveys in Japan. First, the consumer consciousness concerning the purchase of the food has been extracted by the factorial analysis. Second, consumer purchasing behaviors were classified into two types based on factor scores of factorial analysis by cluster analysis. Third, we compared two types and the consumer consciousness that influence purchasing Japanese domestic food was analyzed by structural equation modeling analysis. And last, factors affecting consumers' decision-making in purchasing food were ascertained whether they are based on the consumer

consciousness which influence the purchasing Japanese domestic food by the survey at a store.

A comparison of Pegels and ARIMA models in forecasting sales volume product: case study's SALA ARTIT GARDEN, SURATTHANI Rrovince, Thailand

Kannat Na Bangchang
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand
Email: kannat@mathstat.sci.tu.ac.th

Keywords: Time – series forecasting, Pegels, ARIMA

Abstract: The purpose of this study is comparing the time – series forecasting models based on the Pegels and ARIMA models. The data in this research is real data from the sales productivity of the SALA ARTHIT GARDEN, Ban Nasan district, Suratthani province, Thailand, during the period from January 2003 to December 2014 (amount 144 months). The criterion for comparing in this study is the mean absolute percentage error (MAPE). The study found that the most appropriate in forecasting is the Pegels model in linear trend and addition seasonal, the MAPE is 0.0179. The properties about the estimation are considered.

The time-varying coefficient functional autoregressive model and its application to U.S. treasuries

Meng Xu
School of Economics, Sichuan University, P. R. China
E-mail: meng.xu@stu.scu.edu.cn

Keywords: Sieves; Local regression technique; Smooth structural change; Yield curve.

Abstract: The functional autoregressive (FAR) model which is one of the important special cases among functional data analysis (FDA) has been investigated intensively with many applications, especially to model the autoregressive dynamics of time series data based on infinite dimension. In this paper, we extend the classical constant-coefficient FAR model to the time-varying coefficient FAR model in order to address the non-stationary question by characterizing nonconstancy relationship between functional predictors and functional responses in the model. The estimation of operator of FAR (1) model is considered by the sieves method for dimension reduction, and the local regression technique for the time-varying parameters estimation afterwards. The asymptotic properties of the sieve estimator is obtained to support the modeling procedure. The simulation studies show that the gain by the time-varying FAR modeling is quite substantial, that is, the proposed smooth structural changes can be captured precisely. In the real data analysis,

we apply the yield curves of U.S. government bonds with different maturities for detecting its dynamical patterns which hint the monetary policies shifts smoothly, and the result illustrates consistency between the historical facts and the empirical analysis.

Real-time model calibration in dynamic pricing for high-occupancy toll (HOT) lanes

Cheng Tin Gan, Ph.D.
Department of Civil and Environmental Engineering, Florida International University
E-mail: gana@fiu.edu

Keywords: Dynamic Pricing, High-Occupancy Toll Lanes, Statistical Models, Real-Time Traffic Detector Data, Traffic Simulation

Abstract: High-Occupancy Toll (HOT) lanes are designed to provide more reliable travel times to drivers by adjusting the toll amount as a means to controlling the number of vehicles using the facility. To achieve the objective, HOT lanes are set to maintain a set minimum average speed. Given traffic dynamics, maintaining the minimum speed requires that the toll amount be dynamically set in real time such that it increases or decreases with level of traffic congestion. As such, central to a HOT lane facility is an effective dynamic pricing system that can automatically, consistently, and optimally set the optimal toll amounts in response to real-time traffic conditions. This paper will present a dynamic pricing system that is integrated with two interrelated statistical models: (1) a self-calibration regression model that determines the pricing as a function of historical and real-time detector input including both density and traffic counts, (2) a logit model that estimates the proportion of approaching traffic opting to use the HOT lanes over the general-purpose lanes as a function of the pricing determined by the first model. The system was implemented in a simulation environment and the results show that it is able to meet the minimum speed a high 94% of the time under randomly-generated peak-level traffic. This is significantly higher than the current industry standard of about 85%.

Do we need the constant term in the HAR model for forecasting realized volatilities? Evidence from real sets of data

Hyejin Song
Department of Statistics, Ewha Womans University
E-mail: songhyejin4@gmail.com

Keywords: bias; HAR model; long-memory; Realized volatility; Volatility forecasting

Abstract: No-constant strategy is considered for the heteroscedastic autoregressive (HAR) model of Corsi (2009, A simple approximate long-

memory model of realized volatility). The no-constant model is motivated by smaller biases of its estimated HAR coefficients than those of the constant HAR model. The no-constant model is shown to produce better forecasts than the constant model for 4 real data sets of the realized volatilities of the euro-dollar, the yen-dollar, the pound-euro exchange rates, and the S&P500 index.

A new IHAR model with leverage, called LIHAR model for forecasting realized volatilities

Ji Won Shin
Department of Statistics, Ewha University of Korea.
E-mail: only-thinkme@ewhain.net

Keywords: HAR model, Asymmetry, Realized Volatility forecasting

Abstract: A new strategy is constraining the sum of the HAR coefficients to one, resulting in an integrated model (IHAR). Also we add leverage term to IHAR for asymmetry property of RV. Comparisons of a real data set show substantial advantage of the a new IHAR model with leverage (LIHAR) over the existing HAR and IHAR model. The model is applied for 4 real data sets of RV for 3 US stock price indices (S&P500, NASDAQ, RUSELL2000) and 1 Korean stock price index (KOSPI). The volatilities of the stock price indices are characterized by very persistent long memories and asymmetry. These features are so well-suited for the integrated heteroscedastic autoregressive model with leverage that the LIHAR model produces considerably better out-of-sample forecasts than other models (the HAR model of Corsi 2009) for the 4 real data sets.

Analysis of Korean aircraft departure delay and airport on-time performance due to weather impact

Soyeon Jang
Department of Statistics, Ewha University of Korea.
E-mail: jsoyoun200@gmail.com

Keywords: On time performance, Classification, Gradient boosting machine

Abstract: In this research, we develop statistical models for single flight departure delay and the length of delay time by conditions of weather and airport, especially Jeju International Airport and Gimpo International Airport. Data for this study are collected from the Korea Civil Aviation Development Association and Korea Aviation Meteorological Agency. Logistic regression, randomforest, neural networks, gradient boosting machine, support vector machine and multivariate linear regression are used to investigate the patterns of single aircraft departure on-time performance and delay time. Among them, best methods are selected for building adequate models. From the selected models, it is possible to identify the patterns of aircraft departure

delay. Also, the most significant weather factors to cause the delay are studied.

Classification analysis for unbalanced data

Suyeon Kang
Department of Statistics, Ewha University of Korea.
E-mail: korea92721@naver.com

Keywords: Sampling technique, Asymmetric loss, Misclassification rate, G-mean, Total loss

Abstract: 1. in this paper, we study the classification problem when there are big difference in the proportion of two groups. We call it unbalanced classification problem. In general, it is more difficult to classify the classes accurately in unbalanced data than the balanced data. If we apply the classification methods to the unbalanced data then most of observations are likely to be classified to the bigger group because it can minimize the misclassification loss. However, this misclassification can cause bigger loss in most of real applications.
2. We compare the several classification methods for the unbalanced data using sampling techniques (up and down sampling). We also check the total loss of different classification methods when the asymmetric loss is applied in the simulated and the real data. We use the misclassification rate, g-mean, receiver operating characteristic (ROC) curve and area under the curve (AUC) for the performance comparison.

A comparative study of statistical methods for recurrent survival data analysis

Poster

Sinae Kim
Department of Statistics, Ewha University of Korea.
E-mail: 29anacooll@hanmail.net

Keywords: Recurrent data analysis, survival analysis, Counting process approach, Stratified cox approach, parametric approach

Abstract: Survival data with recurrent events are often collected in the medical studies. Especially, the cancer is the disease that can come back to the same place or another place after a remission. Because the events of a given subject are not independent, some statistical methods for handling the recurrent survival data are proposed. In this paper, we investigate some existing statistical approaches for handling the recurrent survival data - counting process, stratified cox model, parametric model and frailty model. We fit these models to compare the prediction performances by using simulated dataset, and illustrate the methods with real data example.

An application of statistical time series analysis to prediction of slope failure

Tomoaki Imoto
The Institute of Statistical Mathematics, Japan.
E-mail: imoto0923@gmail.com

Keywords: Inverse Gaussian distribution; Surface displacement

Abstract: Landslides have become a major threat and disasters at large scale residence area in urban as well as suburban area or villages. Prediction of slope failure has important roles in public safety against landslides and man made slope failures and the potential of shaping the future of landslides risk management.
The velocity of surface displacement of a slope gives important information on the slope failure and there is a method for predicting the failure time of a slope through the first passage time when the reciprocal of velocity of surface displacement falls to zero. In this paper, we construct a statistical model for the velocity of surface displacement by considering the variation of elements, such as inclination, rainfall, soil and temperature, to be an error. An optimal model is selected by using AIC among ARIMA models and the selected model is related to a random walk. Through the model, we can see the first passage time distribution, or the inverse Gaussian distribution, and predict the slope failure time with probability.

Some contributions to SMC algorithms for option pricing

Deborshee Sen
Department of Statistics and Applied Probability, National University of Singapore.
Email: deborshee.sen@u.nus.edu

Keywords: Option Pricing, SMC, optimal Importance Density, Barrier Options, TARN's.

Abstract: Monte Carlo methods have been used to price options since at least the 1970's. Since the values of assets are often modeled by a stochastic process, computing the price of an option reduces down to computing an expectation with respect to this underlying process. Because of the sequential aspect of the process over time, Sequential Monte Carlo (SMC) methods are a natural tool to apply here. These methods consider a system of interacting particles as opposed to independent particles as in usual Monte Carlo methods. In high dimensional settings when the region of interest is a small part of the sample space, usual Monte Carlo methods can lead to an estimate with high variance. We show that one can achieve significant gains by using SMC methods in such settings by constructing a sequence of artificial target densities over time. In particular, we approximate the optimal importance sampling distribution in the SMC algorithm by using a sequence of weighting functions. This is demonstrated on two examples, barrier options and target

accrual redemption notes (TARN's). This is a joint work with Ajay Jasra and Alexandre Thiery of the National University of Singapore.

Estimation of structural vector autoregression generalized autoregressive conditional heteroscedasticity (SVAR-GARCH) model using independent component analysis (ICA) with applications on high frequency financial data.

Tran Hoang Hai
Department of Statistics and Applied Probability, National University of Singapore, Singapore
Email: a0129183@u.nus.edu

Abstract: In this poster we develop a statistical procedure for the study of the volatility structure in multivariate time series using Structural Vector Autoregression Generalized Autoregressive Conditional Heteroscedasticity (SVAR-GARCH) model. Specifically, using Independent Component Analysis (ICA), we decompose SVAR's error terms into statistically independent time series. Then due to the non-Gaussian nature of ICA components, we suggest using univariate GARCH to model the volatility of each independent component. The structure of volatility transmission along the SVAR model then can be measured using impulse response function to independent component shocks. Empirically, we investigate the the intraday volatility transmission effect along the US Treasury curve, based on high-frequency CME bond futures data.

A combined test for stochastic ordering and crossing survival functions

Hsin-wen Chang
Institute of Statistical Science, Academia Sinica
E-mail: hwchang@stat.sinica.edu.tw

Poster

Keywords: Crossing survival functions; order restricted inference; two-sample problem

Abstract: We develop a test for stochastic ordering, taking into account the possibility of crossing survival functions. By partitioning the parameter space into disjoint hypotheses, our approach is to combine a test that the survival functions do not cross, together with a test of stochastic ordering under the assumption of no crossing. The new procedure is shown via a simulation study to have superior power to the omnibus test in detecting stochastically ordered alternatives, and to operate accurately when survival functions cross.

Online Gaming Data Modeling

Dacheng Chen
dacheng.chen@nus.edu.sg

Abstract: Online gaming is an industry that has been developing very rapidly during recent years. Along with its development, massive amount of data concerning the players' behaviors is generated and collected. One of the most common business models of these games allows players to start the game for free and expect some of them to pay for it after a while. If we take the payment event as an event of interest, then survival model can be applied to this situation. To build a better model, we also take into account the effects from competing risks and potential thresholds within covariates.

Artificial Neural Network versus Linear Models Forecasting Doha Stock Market

Adil Yousif
Department of Math, Stat and Physics, Qatar University
aeyousif@qu.edu.qa

Key words: Neural Network, Time Series, ARIMA, Forecasting, Stock Market

Abstract: The purpose of this Study was to determine the volatility of Doha Stock Market and develop forecasting models. Linear time series models were used and compared to a none linear artificial neural network one namely Multilayer Perceptron Technique. It aims to create models using daily and monthly data collected from Qatar Exchange for the period of January 2007-March 2014. The models generated are for the general index of Qatar Stock Exchange as well as for the several sectors. With the help of these models, the Doha stock market index in general and for various sectors were predicted. The study has made use of various time series techniques to study and analyze the data trend in order to produce appropriate results. It was found that the data acquired a Quadratic trend as it had the lowest MAD and MSD. Quadratic trend model, double exponential smoothing model and ARIMA were applied and it was concluded that ARIMA (2,2) was the most suitable linear model for the daily general index and this matched the results found when the monthly general index was used. However the ANN model was found to be more accurate than time series models.

Thank you for visiting us to Singapore and participating in the IASC-ARS2015 conference!