ims APRM

INSTITUTE OF MATHEMATICAL STATISTICS ASIA PACIFIC RIM MEETING

The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting

June 27-30, 2016

The Chinese University of Hong Kong

Hong Kong

Sponsors:

Croucher Foundation
裘槎基金會

香港中文大學
The Chinese University of Hong Kong

香港中文大學理學院
Faculty of Science, The Chinese University of Hong Kong

New Asia College,
The Chinese University of Hong Kong

Chung Chi College,
The Chinese University of Hong Kong

Institutional Sponsors:

Institute of Mathematical Statistics (IMS)

International Chinese Statistical Association (ICSA)

Bernoulli Society for Mathematical Statistics and Probability

Bernoulli Society (BS)

International Indian Statistical Association (IISA)

Japan Statistical Society (JSS)

Korean Statistical Society (KSS)

Chinese Statistical Association (Taiwan) (CSAT)

香港統計學會
Hong Kong Statistical Society

Hong Kong Statistical Society (HKSS)

Chinese Society of Probability and Statistics (CPSS)

ims APRM

INSTITUTE OF MATHEMATICAL STATISTICS ASIA PACIFIC RIM MEETING

The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting

June 27-30, 2016

The Chinese University of Hong Kong

Hong Kong

Sponsors:

Croucher Foundation 裘槎基金會

香港中文大學 The Chinese University of Hong Kong

香港中文大學理學院 Faculty of Science, The Chinese University of Hong Kong

New Asia College, The Chinese University of Hong Kong

Chung Chi College, The Chinese University of Hong Kong

Institutional Sponsors:

Institute of Mathematical Statistics (IMS)

International Chinese Statistical Association (ICSA)

Bernoulli Society (BS)

International Indian Statistical Association (IISA)

Japan Statistical Society (JSS)

Korean Statistical Society (KSS)

Chinese Statistical Association (Taiwan) (CSAT)

Hong Kong Statistical Society (HKSS)

Chinese Society of Probability and Statistics (CPSS)

Scientific Program Committee

Ming-Yen Cheng (Co-chair), National Taiwan University
Xuming He (Co-chair), University of Michigan
Makoto Aoshima, University of Tsukuba
Amarjit Budhiraja, The University of North Carolina at Chapel Hill
Probal Chaudhuri, Indian Statistical Institute
Woncheol Jang, Seoul National University
Wai Keung Li, The University of Hong Kong
Judith Rousseau, Université Paris Dauphine
Qi-Man Shao, The Chinese University of Hong Kong
Niansheng Tang, Yunnan University
Alan Welsh, Australian National University
Jin-Ting Zhang, National University of Singapore

Local Organizing Committee

Qi-Man Shao (Chair), The Chinese University of Hong Kong
Siu Hung Cheung (Secretary), The Chinese University of Hong Kong
Ping Shing Chan, The Chinese University of Hong Kong
Xiaodan Fan, The Chinese University of Hong Kong
Bing-Yi Jing, The Hong Kong University of Science and Technology
Yuanyuan Lin, The Chinese University of Hong Kong
Tony Sit, The Chinese University of Hong Kong
Xin Yuan Song, The Chinese University of Hong Kong
Yingying Wei, The Chinese University of Hong Kong
Hoi Ying Wong, The Chinese University of Hong Kong
Phillip Sheung Chi Yam, The Chinese University of Hong Kong
Jian-Feng Yao, The University of Hong Kong
Chun Yip Yau, The Chinese University of Hong Kong
Lixing Zhu, Hong Kong Baptist University

Sponsor Information

Sponsors

Croucher Foundation 裘槎基金會	Croucher Foundation	www.croucher.org.hk
香港中文大學 The Chinese University of Hong Kong	The Chinese University of Hong Kong	www.cuhk.edu.hk
香港中文大學理學院 Faculty of Science, The Chinese University of Hong Kong	Faculty of Science, The Chinese University of Hong Kong	www.cuhk.edu.hk/sci
NEW ASIA COLLEGE	New Asia College, The Chinese University of Hong Kong	www.na.cuhk.edu.hk
	Chung Chi College, The Chinese University of Hong Kong	www.ccc.cuhk.edu.hk

Institutional Sponsors

Institute of Mathematical Statistics (IMS)	http://imstat.org
International Chinese Statistical Association (ICSA)	www.icsa.org
Bernoulli Society (BS)	www.bernoulli-society.org
International Indian Statistical Association (IISA)	www.intindstat.org
Japan Statistical Society (JSS)	www.jss.gr.jp
Korean Statistical Society (KSS)	www.kss.or.kr
Chinese Statistical Association (Taiwan) (CSAT)	www.stat.org.tw
Hong Kong Statistical Society (HKSS)	www.hkss.org.hk
Chinese Society of Probability and Statistics (CPSS)	http://math0.bnu.edu.cn/statprob

Program Schedule

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Sun, June 26						
18:00-20:00	Registration (YIA G/F)					
18:00-20:30	*Welcome Reception (YIA G/F)*					
Mon, June 27						
08:00-17:00	Registration and Enquiry (YIA G/F)					
09:00-09:40	**Opening Ceremony (YIA G/F LT1)**					
09:40-10:00	*Coffee Break (YIA G/F & 2/F)*					
10:00-11:00	**Plenary Lecture 1 (YIA G/F LT1)** **Understanding Importance Sampling** Plenary Speaker: Persi Diaconis (Stanford University) Chair: Qi-Man Shao (The Chinese University of Hong Kong)					
11:10-12:10	**Plenary Lecture 2 (YIA G/F LT1)** **Low Rank Structure in Highly Multivariate Models** Plenary Speaker: Iain Johnstone (Stanford University) Chair: Ming-Yen Cheng (National Taiwan University)					
12:10-13:30	*Lunch (Chung Chi Tang Student Canteen)*					
13:30-15:10	**YIA G/F LT2**	**YIA G/F LT3**	**YIA 2/F LT4**	**YIA 2/F LT5**	**YIA 2/F LT6**	**YIA 2/F LT7**
	DL07: Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics	**DL09:** Recent Advances in Covariance Estimation	**IP01:** Recent Advances on Analysis of High-Dimensional Data	**IP34:** Robust Statistical Methods	**IP37:** Multivariate Smoothing	**IP21:** Complex Data Analysis
15:10-15:30	*Coffee Break (YIA G/F & 2/F)*					
15:30-17:10	**DL10:** Statistical Analysis for Social Network Data	**DL17:** Sparse Learning from Covariance	**IP07:** Big Data Integration	**IP12:** Longitudinal Data Analysis	**IP15:** Limit Theorems in Stochastic Analysis	**IP20:** Nonparametric Inference in Econometric Models

DL: Distinguished Lectures • IP: Invited Paper Sessions • TCP: Topic Contributed Paper Sessions • CP: Contributed Paper Sessions

YIA 2/F LT8	YIA 2/F LT9	YIA 2/F 201	YIA 4/F 403	YIA 4/F 405	YIA 5/F 505
IP51: Complex Data Analysis: Theory and Methods	**TCP15:** Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques II: Robust SEM, Multilevel SEM, Meta-Analytic SEM, Latent Profile Analysis, and Cross-Classified SEM	**TCP16:** Recent Advances and Challenges in Analysis of Complex Biomedical Data	**TCP17:** Recent Advances in Object Data Analysis	**CP01:** Machine Learning Session 1	
IP57: Recent Advances and Challenges of Big Data Inference with Complex Structures	**TCP12:** Analysis of Reliability and Survival Data	**TCP13:** Theory and Applications of Tensor Variate Data Analysis	**TCP14:** Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques I: Measurement Invariance and Latent Growth Models	**CP02:** Machine Learning Session 2	

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Tue, June 28						
08:00-17:00	Registration and Enquiry (YIA G/F)					
08:30-10:10	**YIA G/F LT2**	**YIA G/F LT3**	**YIA 2/F LT4**	**YIA 2/F LT5**	**YIA 2/F LT6**	**YIA 2/F LT7**
	DL13: Recent Development of Statistical Inference Under Dimension Reduction Structure	**IP60:** Special Session in Memory of Peter Hall	**IP16:** Models and Inference for Complex Extreme Events	**IP28:** Asymptotic Theory	**IP32:** Distributions on Convex Cones and Sets of Permutations	**IP35:** Survey Sampling and Related Methodologies
10:10-10:30	*Coffee Break (YIA G/F & 2/F)*					
10:30-12:10	**DL02:** Building Bridges: New Bayesian Insights into Old Problems	**IP02:** Modern Developments in Multivariate Data	**IP03:** Statistical Analysis of Complex Correlated Data	**IP44:** Stein's Method and Random Graphs	**IP18:** New Frontiers of Longitudinal/ Clustered Data Analysis	**IP38:** Dimension Reduction and Variable Selection for High-Dimensional Data
12:10-13:30	*Lunch (Chung Chi Tang Student Canteen)*					
13:00-13:45	**Poster Session (YIA G/F)**					
13:30-15:10	**DL12:** Nonstationary Time Series: Past, Present and Beyond	**IP05:** Recent Advances in Large-Scale Inference	**IP19:** Recent Advances in Bayesian Functional Data Analysis	**IP23:** Recent Advances in Design of Experiments	**IP25:** Biostatistics and Biomedical Statistics	**IP30:** Recent Advances on Algebraic Methods in Statistics
15:10-15:30	*Coffee Break (YIA G/F & 2/F)*					
15:30-17:10	**DL06:** Model Selection in High-Dimensional Regression and Graphical Models and Its Applications	**IP06:** Recent Advances in Big Data Inference	**IP27:** Time Series and Econometrics	**IP29:** Education of Data Science and Statistics	**IP24:** Recent Advances in Functional Data Analysis	**IP31:** Recent Advances in Non- and Semi-Parametric Inference
Conference Banquet Reception & Dinner (Science Park, ClubONE on the Park) [tickets required] Address: Shop061-066, G/F, Building 12W, No.12 Science Park West Avenue, Hong Kong Science Park, Shatin, N.T. *- Coaches will start departing at the entrance of YIA during 17:15-18:00.* *-Transportation from ClubONE on the Park to University MTR Station will be provided after the banquet.*						
17:30-19:00	*Conference Banquet Reception (Science Park, ClubONE on the Park)*					
19:00-21:30	*Conference Banquet Dinner (Science Park, ClubONE on the Park)*					

DL: Distinguished Lectures • IP: Invited Paper Sessions • TCP: Topic Contributed Paper Sessions • CP: Contributed Paper Sessions

YIA 2/F LT8	**YIA 2/F LT9**	**YIA 2/F 201**	**YIA 4/F 403**	**YIA 4/F 405**	**YIA 5/F 505**
TCP03: Statistics and Computing for Complex Dependent Systems	**TCP08:** Stochastic Processes and Related Topics	**TCP23**: Recent Advances in Time-to-Event Analysis	**TCP06**: Geometric Aspects of Statistical Methods	**CP14:** Data Order Structure Session 1	**DL11:** Design of Experiments
IP54: Recent Advances in Analysis of Complex Data	**TCP18:** Recent Advances in Statistical Genetics	**TCP19:** Recent Developments in Model Checking	**TCP10:** Bayesian Modeling with Complex Survey Data	**CP03:** Machine Learning Session 3	
IP56: Modern Statistical Methods for Complex Data	**TCP21:** Confidence Interval in Some Models and High Dimensional Data Analysis	**TCP22:** Numerical Computation Methods for Multivariate Distributions and Parameter Estimation	**TCP20:** Statistical Analysis of Survival and Functional Data	**CP04:** Machine Learning Session 4	
TCP24: Recent Topics in Medical Statistics	**TCP27:** Recent Developments in Biostatistics and Spatial Statistics	**TCP28:** Recent Developments in Multiple Comparison Procedures	**TCP31:** New Statistical Methods for Graphs and Networks	**CP15:** Data Order Structure Session 2	

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Wed, June 29						
08:30-10:10	YIA G/F LT2	YIA G/F LT3	YIA 2/F LT4	YIA 2/F LT5	YIA 2/F LT6	YIA 2/F LT7
	DL03: Fusion Learning - Fusing Inferences from Diverse Sources	**IP04:** Statistical Analysis of Dynamic Models	**IP08:** Statistical Advances of Image Analysis and Spatial Statistics	**IP42:** Recent Advances in Analysis of Complex Observational Data	**IP10:** Development in Statistical Methods for the Analysis of Events	**IP13:** Stochastic Interacting Systems
10:10-10:30	*Coffee Break (YIA G/F & 2/F)*					
10:30-12:10	**DL08:** Heterogeneity in Large-Scale Data, with Connections to Causal Inference	**DL16:** Variational Inference	**IP11:** Nonparametric Modelling and High Dimensional Data Analysis	**IP22:** Recent Advances in Complex Data Analysis	**IP43:** Statistical Methodology for Biomedical Sciences	**IP52:** Challenges and Recent Advances in Methods for Missing Data Problems
12:10-13:30	*Lunch (Chung Chi Tang Student Canteen)*					
13:45-17:00	*Excursion [tickets required]* *Route 1: Hong Kong City Tour (The Peak, Repulse Bay, Stanley and Murray House)* *Route 2: Hong Kong Cultural Tour (Ngong Ping Piazza, Tian Tan Buddha, Po Lin Monastery and Wisdom Path)* *- Coaches will start departing at the entrance of YIA during 13:45-14:00*					

DL: Distinguished Lectures • IP: Invited Paper Sessions • TCP: Topic Contributed Paper Sessions • CP: Contributed Paper Sessions

YIA 2/F LT8	YIA 2/F LT9	YIA 2/F 201	YIA 4/F 403	YIA 4/F 405	YIA 5/F 505
IP55: New Developments in Statistical Genomics	**TCP01:** Recent Advances in Bayesian Nonparametrics	**TCP04:** Variable Selection and Applications of Semiparametric Models	**TCP32:** Model Selection and Hypothesis Testing	**CP06:** Bayesian Session 1	**IP41:** New Developments in High-Dimensional Spatial and Spatio-Temporal Modeling
TCP11: New Statistical Methods for Genetic Data Analysis	**TCP26:** Statistical Issues in Analyzing Bioinformatics Data	**TCP29:** Recent Advances on Random Processes and Related Problems	**TCP30:** Statistical Inference Under Model Uncertainties	**CP08:** Probability Session 1	

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Thu, June 30						
08:30-10:10	**YIA G/F LT2**	**YIA G/F LT3**	**YIA 2/F LT4**	**YIA 2/F LT5**	**YIA 2/F LT6**	**YIA 2/F LT7**
	DL01: Recent Advances in Machine Learning for Personalized Medicine	**IP26:** Random Matrices and High-Dimensional Statistics	**IP09:** Adaptive Randomization in Clinical Trials	**IP40:** Recent Developments in the Analysis of High-Dimensional Time Series with Nonstationarities	**IP14:** Network Models: Theory and Methods	**IP36:** Statistical Inferences for Complex Data
10:10-10:30	*Coffee Break (YIA G/F & 2/F)*					
10:30-12:10	**DL05:** Particle Representations for Measure-Valued Processes and Stochastic Partial Differential Equations		**IP45:** Random Fields: Theory and Applications	**IP49:** Emerging Statistical Methods in Big Data Analytics	**IP47:** Emerging Nonparametric Methods for Financial Data	**IP59:** Advances in Analysis of Complex High-Dimensional Data
12:10-13:30	*Lunch (Chung Chi Tang Student Canteen)*					
13:00-13:45	**Poster Session (YIA G/F)**					
13:30-15:10	**DL04:** From Cells to Populations: Modeling and Inference for Genomic Data	**DL15:** Random Networks	**IP48:** Recent Advances in Lifetime Data Analysis	**IP46:** Recent Developments in Survival Analysis and Personalized Medicine	**IP53:** Some Recent Developments in High-Frequency Financial Econometrics	**IP33:** Analysis of Spatial and Spatio-Temporal Data
15:10-15:30	*Coffee Break (YIA G/F & 2/F)*					
15:30-17:10	**DL14:** Statistical Inference for Stochastic Processes: Asymptotic Theory and Implementation	**IP17:** Recent Advances and Trends in Time Series Analysis	**IP39:** Advances in Statistical Inference for Multivariate Response Data		**IP50:** Semiparametric Statistical Methods for Complex Problems	**IP58:** Stochastic Partial Differential Equations

DL: Distinguished Lectures • IP: Invited Paper Sessions • TCP: Topic Contributed Paper Sessions • CP: Contributed Paper Sessions

YIA 2/F LT8	YIA 2/F LT9	YIA 2/F 201	YIA 4/F 403	YIA 4/F 405	YIA 5/F 505
TCP02: Advanced Modeling of Large-Scale Dependent Data	**TCP25:** Recent Developments in Large and Complex Data Analysis	**TCP34:** Statistical Modeling and Its Applications	**TCP35:** Statistical Modeling in Economics and Finance	**CP09:** Probability Session 2	
TCP09: Recent Advances in Biomarker Evaluation and Risk Prediction	**TCP33:** Advanced Bayesian Modeling	**TCP07:** Recent Advances in High Dimensional Estimation Theory	**TCP05:** High-Dimensional Data Analyses with Application in Biomedical Studies	**CP11:** Big Data/ Network Session 1	
	CP12: Big Data/ Network Session 2	**CP10:** Probability Session 3	**CP16:** Data Order Structure Session 3	**CP17:** Data Order Structure Session 4	
CP07: Bayesian Session 2	**CP05:** Machine Learning Session 5	**CP13:** Big Data/ Network Session 3	**CP18:** Data Order Structure Session 5	**CP19:** Data Order Structure Session 6	

Oral Presentations

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Mon, June 27

Time	Session	Session Title	Venue
10:00-11:00			
	PL01	Understanding Importance Sampling	YIA G/F LT1
11:10-12:10			
	PL02	Low Rank Structure in Highly Multivariate Models	YIA G/F LT1
13:30-15:10			
	DL07	Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics	YIA G/F LT2
	DL09	Recent Advances in Covariance Estimation	YIA G/F LT3
	IP01	Recent Advances on Analysis of High-Dimensional Data	YIA 2/F LT4
	IP34	Robust Statistical Methods	YIA 2/F LT5
	IP37	Multivariate Smoothing	YIA 2/F LT6
	IP21	Complex Data Analysis	YIA 2/F LT7
	IP51	Complex Data Analysis: Theory and Methods	YIA 2/F LT8
	TCP15	Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques II: Robust SEM, Multilevel SEM, Meta-Analytic SEM, Latent Profile Analysis, and Cross-Classified SEM	YIA 2/F LT9
	TCP16	Recent Advances and Challenges in Analysis of Complex Biomedical Data	YIA 2/F 201
	TCP17	Recent Advances in Object Data Analysis	YIA 4/F 403
	CP01	Machine Learning Session 1	YIA 4/F 405

Mon, June 27

Time	Session	Session Title	Venue
15:30-17:10			
	DL10	Statistical Analysis for Social Network Data	YIA G/F LT2
	DL17	Sparse Learning from Covariance	YIA G/F LT3
	IP07	Big Data Integration	YIA 2/F LT4
	IP12	Longitudinal Data Analysis	YIA 2/F LT5
	IP15	Limit Theorems in Stochastic Analysis	YIA 2/F LT6
	IP20	Nonparametric Inference in Econometric Models	YIA 2/F LT7
	IP57	Recent Advances and Challenges of Big Data Inference with Complex Structures	YIA 2/F LT8
	TCP12	Analysis of Reliability and Survival Data	YIA 2/F LT9
	TCP13	Theory and Applications of Tensor Variate Data Analysis	YIA 2/F 201
	TCP14	Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques I: Measurement Invariance and Latent Growth Models	YIA 4/F 403
	CP02	Machine Learning Session 2	YIA 4/F 405

Tue, June 28

	Session	Session Title	Venue
08:30-10:10			
	DL13	Recent Development of Statistical Inference Under Dimension Reduction Structure	YIA G/F LT2
	IP60	Special Session in Memory of Peter Hall	YIA G/F LT3
	IP16	Models and Inference for Complex Extreme Events	YIA 2/F LT4
	IP28	Asymptotic Theory	YIA 2/F LT5
	IP32	Distributions on Convex Cones and Sets of Permutations	YIA 2/F LT6
	IP35	Survey Sampling and Related Methodologies	YIA 2/F LT7
	TCP03	Statistics and Computing for Complex Dependent Systems	YIA 2/F LT8
	TCP08	Stochastic Processes and Related Topics	YIA 2/F LT9
	TCP23	Recent Advances in Time-to-Event Analysis	YIA 2/F 201
	TCP06	Geometric Aspects of Statistical Methods	YIA 4/F 403
	CP14	Data Order Structure Session 1	YIA 4/F 405
	DL11	Design of Experiments	YIA 5/F 505
10:30-12:10			
	DL02	Building Bridges: New Bayesian Insights into Old Problems	YIA G/F LT2
	IP02	Modern Developments in Multivariate Data	YIA G/F LT3
	IP03	Statistical Analysis of Complex Correlated Data	YIA 2/F LT4
	IP44	Stein's Method and Random Graphs	YIA 2/F LT5
	IP18	New Frontiers of Longitudinal/Clustered Data Analysis	YIA 2/F LT6
	IP38	Dimension Reduction and Variable Selection for High-Dimensional Data	YIA 2/F LT7
	IP54	Recent Advances in Analysis of Complex Data	YIA 2/F LT8
	TCP18	Recent Advances in Statistical Genetics	YIA 2/F LT9
	TCP19	Recent Developments in Model Checking	YIA 2/F 201
	TCP10	Bayesian Modeling with Complex Survey Data	YIA 4/F 403
	CP03	Machine Learning Session 3	YIA 4/F 405

Tue, June 28

Time	Session	Session Title	Venue
13:30-15:10			
	DL12	Nonstationary Time Series: Past, Present and Beyond	YIA G/F LT2
	IP05	Recent Advances in Large-Scale Inference	YIA G/F LT3
	IP19	Recent Advances in Bayesian Functional Data Analysis	YIA 2/F LT4
	IP23	Recent Advances in Design of Experiments	YIA 2/F LT5
	IP25	Biostatistics and Biomedical Statistics	YIA 2/F LT6
	IP30	Recent Advances on Algebraic Methods in Statistics	YIA 2/F LT7
	IP56	Modern Statistical Methods for Complex Data	YIA 2/F LT8
	TCP21	Confidence Interval in Some Models and High Dimensional Data Analysis	YIA 2/F LT9
	TCP22	Numerical Computation Methods for Multivariate Distributions and Parameter Estimation	YIA 2/F 201
	TCP20	Statistical Analysis of Survival and Functional Data	YIA 4/F 403
	CP04	Machine Learning Session 4	YIA 4/F 405
15:30-17:10			
	DL06	Model Selection in High-Dimensional Regression and Graphical Models and Its Applications	YIA G/F LT2
	IP06	Recent Advances in Big Data Inference	YIA G/F LT3
	IP27	Time Series and Econometrics	YIA 2/F LT4
	IP29	Education of Data Science and Statistics	YIA 2/F LT5
	IP24	Recent Advances in Functional Data Analysis	YIA 2/F LT6
	IP31	Recent Advances in Non- and Semi-Parametric Inference	YIA 2/F LT7
	TCP24	Recent Topics in Medical Statistics	YIA 2/F LT8
	TCP27	Recent Developments in Biostatistics and Spatial Statistics	YIA 2/F LT9
	TCP28	Recent Developments in Multiple Comparison Procedures	YIA 2/F 201
	TCP31	New Statistical Methods for Graphs and Networks	YIA 4/F 403
	CP15	Data Order Structure Session 2	YIA 4/F 405

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Wed, June 29

	Session	Session Title	Venue
08:30-10:10			
	DL03	Fusion Learning - Fusing Inferences from Diverse Sources	YIA G/F LT2
	IP04	Statistical Analysis of Dynamic Models	YIA G/F LT3
	IP08	Statistical Advances of Image Analysis and Spatial Statistics	YIA 2/F LT4
	IP42	Recent Advances in Analysis of Complex Observational Data	YIA 2/F LT5
	IP10	Development in Statistical Methods for the Analysis of Events	YIA 2/F LT6
	IP13	Stochastic Interacting Systems	YIA 2/F LT7
	IP55	New Developments in Statistical Genomics	YIA 2/F LT8
	TCP01	Recent Advances in Bayesian Nonparametrics	YIA 2/F LT9
	TCP04	Variable Selection and Applications of Semiparametric Models	YIA 2/F 201
	TCP32	Model Selection and Hypothesis Testing	YIA 4/F 403
	CP06	Bayesian Session 1	YIA 4/F 405
	IP41	New Developments in High-Dimensional Spatial and Spatio-Temporal Modeling	YIA 5/F 505

Wed, June 29

Time	Session	Session Title	Venue
10:30-12:10			
	DL08	Heterogeneity in Large-Scale Data, with Connections to Causal Inference	YIA G/F LT2
	DL16	Variational Inference	YIA G/F LT3
	IP11	Nonparametric Modelling and High Dimensional Data Analysis	YIA 2/F LT4
	IP22	Recent Advances in Complex Data Analysis	YIA 2/F LT5
	IP43	Statistical Methodology for Biomedical Sciences	YIA 2/F LT6
	IP52	Challenges and Recent Advances in Methods for Missing Data Problems	YIA 2/F LT7
	TCP11	New Statistical Methods for Genetic Data Analysis	YIA 2/F LT8
	TCP26	Statistical Issues in Analyzing Bioinformatics Data	YIA 2/F LT9
	TCP29	Recent Advances on Random Processes and Related Problems	YIA 2/F 201
	TCP30	Statistical Inference Under Model Uncertainties	YIA 4/F 403
	CP08	Probability Session 1	YIA 4/F 405

Venue: Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Thu, June 30

Time	Session	Session Title	Venue
08:30-10:10			
	DL01	Recent Advances in Machine Learning for Personalized Medicine	YIA G/F LT2
	IP26	Random Matrices and High-Dimensional Statistics	YIA G/F LT3
	IP09	Adaptive Randomization in Clinical Trials	YIA 2/F LT4
	IP40	Recent Developments in the Analysis of High-Dimensional Time Series with Nonstationarities	YIA 2/F LT5
	IP14	Network Models: Theory and Methods	YIA 2/F LT6
	IP36	Statistical Inferences for Complex Data	YIA 2/F LT7
	TCP02	Advanced Modeling of Large-Scale Dependent Data	YIA 2/F LT8
	TCP25	Recent Developments in Large and Complex Data Analysis	YIA 2/F LT9
	TCP34	Statistical Modeling and Its Applications	YIA 2/F 201
	TCP35	Statistical Modeling in Economics and Finance	YIA 4/F 403
	CP09	Probability Session 2	YIA 4/F 405
10:30-12:10			
	DL05	Particle Representations for Measure-Valued Processes and Stochastic Partial Differential Equations	YIA G/F LT2
	IP45	Random Fields: Theory and Applications	YIA 2/F LT4
	IP49	Emerging Statistical Methods in Big Data Analytics	YIA 2/F LT5
	IP47	Emerging Nonparametric Methods for Financial Data	YIA 2/F LT6
	IP59	Advances in Analysis of Complex High-Dimensional Data	YIA 2/F LT7
	TCP09	Recent Advances in Biomarker Evaluation and Risk Prediction	YIA 2/F LT8
	TCP33	Advanced Bayesian Modeling	YIA 2/F LT9
	TCP07	Recent Advances in High Dimensional Estimation Theory	YIA 2/F 201
	TCP05	High-Dimensional Data Analyses with Application in Biomedical Studies	YIA 4/F 403
	CP11	Big Data/Network Session 1	YIA 4/F 405

Thu, June 30

	Session	Session Title	Venue
13:30-15:10			
	DL04	From Cells to Populations: Modeling and Inference for Genomic Data	YIA G/F LT2
	DL15	Random Networks	YIA G/F LT3
	IP48	Recent Advances in Lifetime Data Analysis	YIA 2/F LT4
	IP46	Recent Developments in Survival Analysis and Personalized Medicine	YIA 2/F LT5
	IP53	Some Recent Developments in High-Frequency Financial Econometrics	YIA 2/F LT6
	IP33	Analysis of Spatial and Spatio-Temporal Data	YIA 2/F LT7
	CP12	Big Data/Network Session 2	YIA 2/F LT9
	CP10	Probability Session 3	YIA 2/F 201
	CP16	Data Order Structure Session 3	YIA 4/F 403
	CP17	Data Order Structure Session 4	YIA 4/F 405
15:30-17:10			
	DL14	Statistical Inference for Stochastic Processes: Asymptotic Theory and Implementation	YIA G/F LT2
	IP17	Recent Advances and Trends in Time Series Analysis	YIA G/F LT3
	IP39	Advances in Statistical Inference for Multivariate Response Data	YIA 2/F LT4
	IP50	Semiparametric Statistical Methods for Complex Problems	YIA 2/F LT6
	IP58	Stochastic Partial Differential Equations	YIA 2/F LT7
	CP07	Bayesian Session 2	YIA 2/F LT8
	CP05	Machine Learning Session 5	YIA 2/F LT9
	CP13	Big Data/Network Session 3	YIA 2/F 201
	CP18	Data Order Structure Session 5	YIA 4/F 403
	CP19	Data Order Structure Session 6	YIA 4/F 405

Poster Presentations

Venue: G/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Tue-Thu, June 28-30

PS02	Birth Cohort Effect in Japan - Automatic Detection and Statistical Evaluation
PS03	High Dimensional LASSO Variable Selection Under Strongly-correlated Covariates
PS04	Growth Curve Model with Nonparametric Baselines and Its Statistical Inference
PS05	Evaluating Standard Errors of Total Heritability Estimate in Genome-wide Association Studies Based on Summary Statistics Alone
PS06	Estimating Regression Coefficients Including Nuisance Baseline and Its Applications
PS07	Joint Inference for a GLMM Model with a NLME Covariate Model Subject to Left Censoring and Measurement Error, with Application to AIDS Studies
PS09	A New Distribution to Describe Big Data
PS10	Nonlinear Operator Estimation with Bayes Sieve Estimator
PS11	On the Spectral Distribution of Hayashi's Estimator for High Dimensional Stock Price Process
PS12	Recent Advances in Approximate Solution for Stochastic Differential Delay Equation
PS13	Unified Tree-structured Non-crossing Quantile Regression Model

Poster Session

Core time:	Tue, June 28, 13:00-13:45, 45 minutes Thu, June 30, 13:00-13:45, 45 minutes

Session Presenters' Guide

(Oral and Poster Presentations)

Guide for Oral Session Presenters Session Room

Location:	Each session room
Operation Hour:	Mon, June 27, 08:30-17:10 Tue, June 28, 08:15-17:10 Wed, June 29, 08:15-12:10 Thu, June 30, 08:15-17:10
Support Staff:	1 staff per session room

Guideline:

1. Each session room will be equipped with a computer and a projector. The operating system for the session room computer will be Microsoft Windows 7, with PowerPoint 2010 and PDF reader installed.
2. Each session room computer will be equipped with USB ports to read USB flash memory stick. The presenter must load the presentation file by him/herself onto the computer through these media.
3. The presenters should load the presentation data onto the session room computer 15 minutes prior to the start of the session.

Network Environment at the Venue

Connection Type:	Wireless
Wi-Fi SSID:	CUguest
Login User ID:	4imsaprm@conference.cuhk.edu.hk
Login Password:	cuhk2016
Start Valid Time:	Sun, June 26
End Valid Time:	Thu, June 30

Guide for Poster Session Presenters

• Poster Board

Location:	G/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong
Operation Hour:	From 13:00 on June 28 to 17:00 on June 30
Poster Size:	For best viewing, the suggested poster size is 38 inches (96.5cm in width) x 48 inches (122cm in height)
Mounting:	Push pins for mounting the posters will be available at the poster display area
Set up hour:	08:30-12:00 on June 28
Removal :	before 17:00 on June 30
Notice:	Posters remain on the board after 17:00 on June 30 will be removed and disposed by the secretariat office.

• Poster Session

Core time:	Tue, June 28, 13:00-13:45, 45 minutes Thu, June 30, 13:00-13:45, 45 minutes

Conference Events

SUN, JUNE 26

Welcome Reception

Date: Sun, June 26
Time: 18:00-20:30
Venue: G/F, Yasumoto International Academic Park (YIA),
The Chinese University of Hong Kong

MON, JUNE 27

Opening Ceremony

Date: Mon, June 27
Time: 9:00-9:40
Venue: Lecture Theatre 1, Yasumoto International Academic Park (YIA),
The Chinese University of Hong Kong

MON-THU, JUNE 27-30

Lunch

Lunch will be provided at Chung Chi Tang Student Canteen, Chung Chi College during 12:10-13:30 from June 27 to June 30.

Coffee Break

Coffee break will be provided at G/F and 2/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong from June 27 to June 30.

TUE, JUNE 28

Conference Banquet [tickets required]

Date: Tue, June 28
Time: Reception starts from 17:30; Dinner starts from 19:00
Venue: ClubONE on the Park, Science Park,
Shop061-066, G/F, Building 12W,
No.12 Science Park West Avenue,
Hong Kong Science Park, Sha Tin, N.T.

(Coaches will start departing at the entrance of YIA during 17:15-18:00. Transportation from ClubONE on the Park to University MTR Station will be provided after the banquet.)

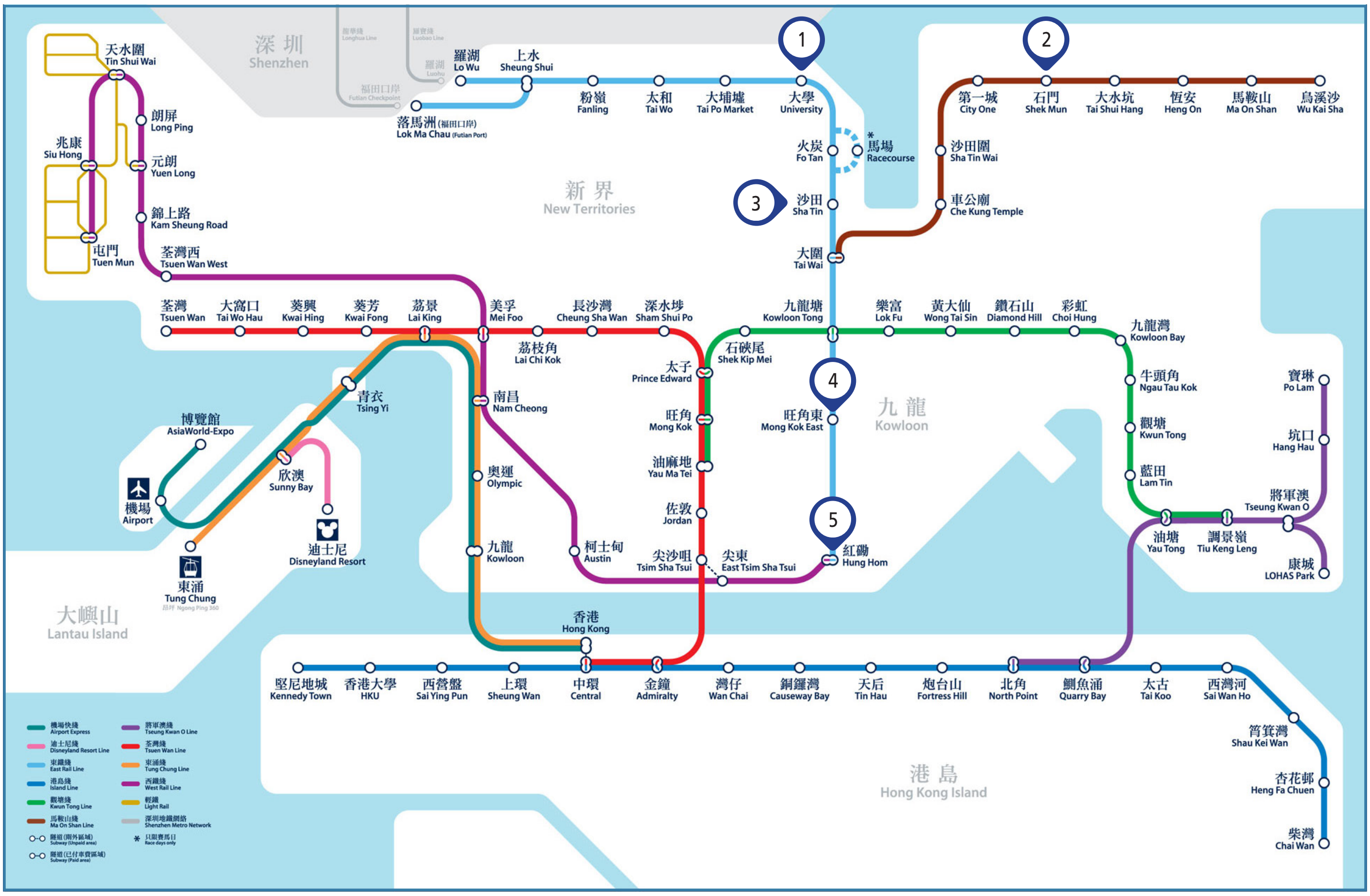
WED, JUNE 29

Excursion [tickets required]

Date: Wed, June 29
Time: 14:00-17:00
Route 1: Hong Kong City Tour (The Peak, Repulse Bay, Stanley and Murray House)
Route 2: Hong Kong Cultural Tour (Ngong Ping Piazza, Tian Tan Buddha, Po Lin Monastery and Wisdom Path)

(Coaches will start departing at the entrance of YIA during 13:45-14:00)

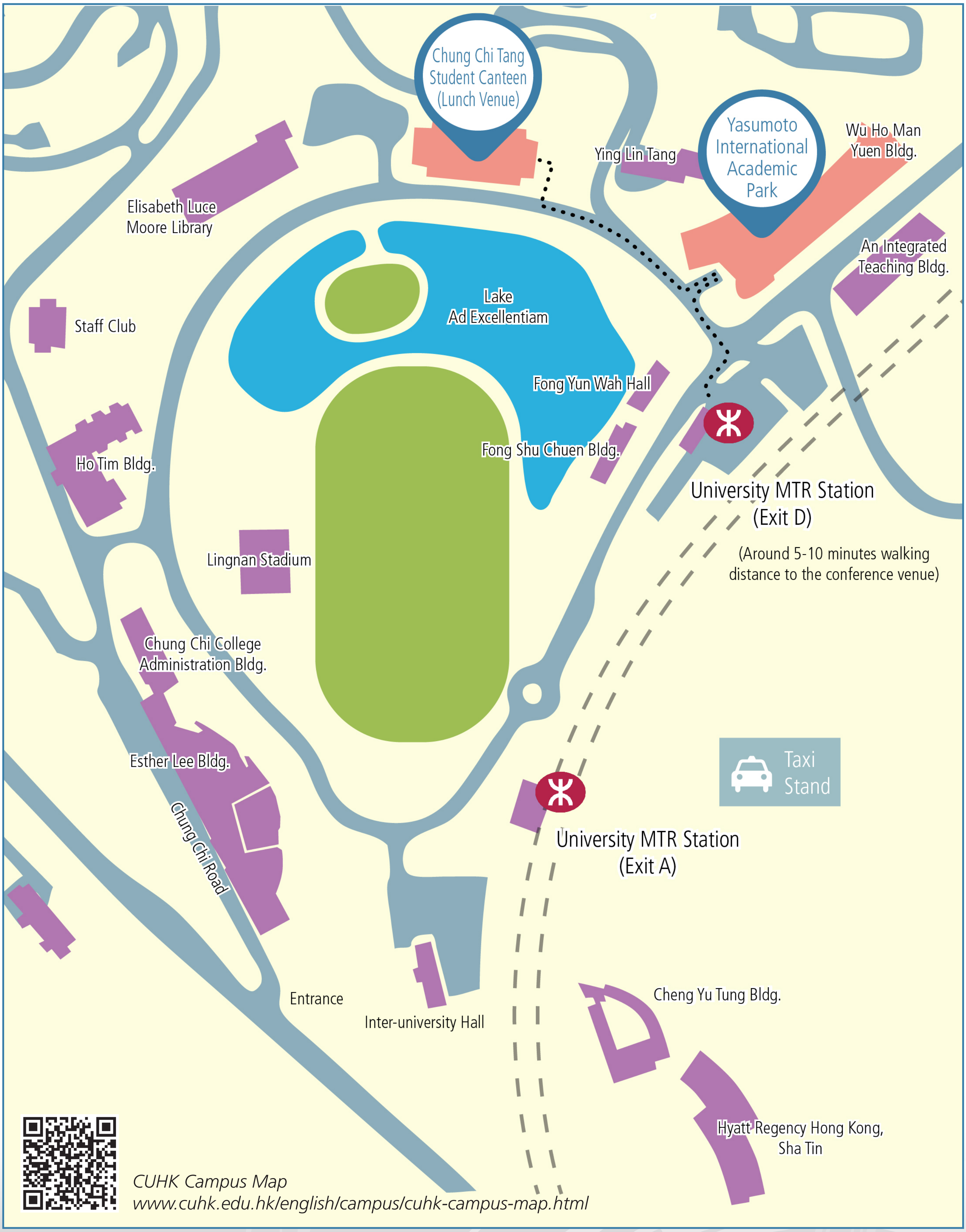
MTR System Map



1	University Station	(i) **The Chinese University of Hong Kong** (ii) Hyatt Regency Hong Kong, Sha Tin
2	Shek Mun Station	Courtyard by Marriott Hong Kong, Sha Tin
3	Sha Tin Station	Royal Park Hotel
4	*Mong Kok East Station	Royal Plaza Hotel
5	Hung Hom Station	Harbour Plaza Metropolis

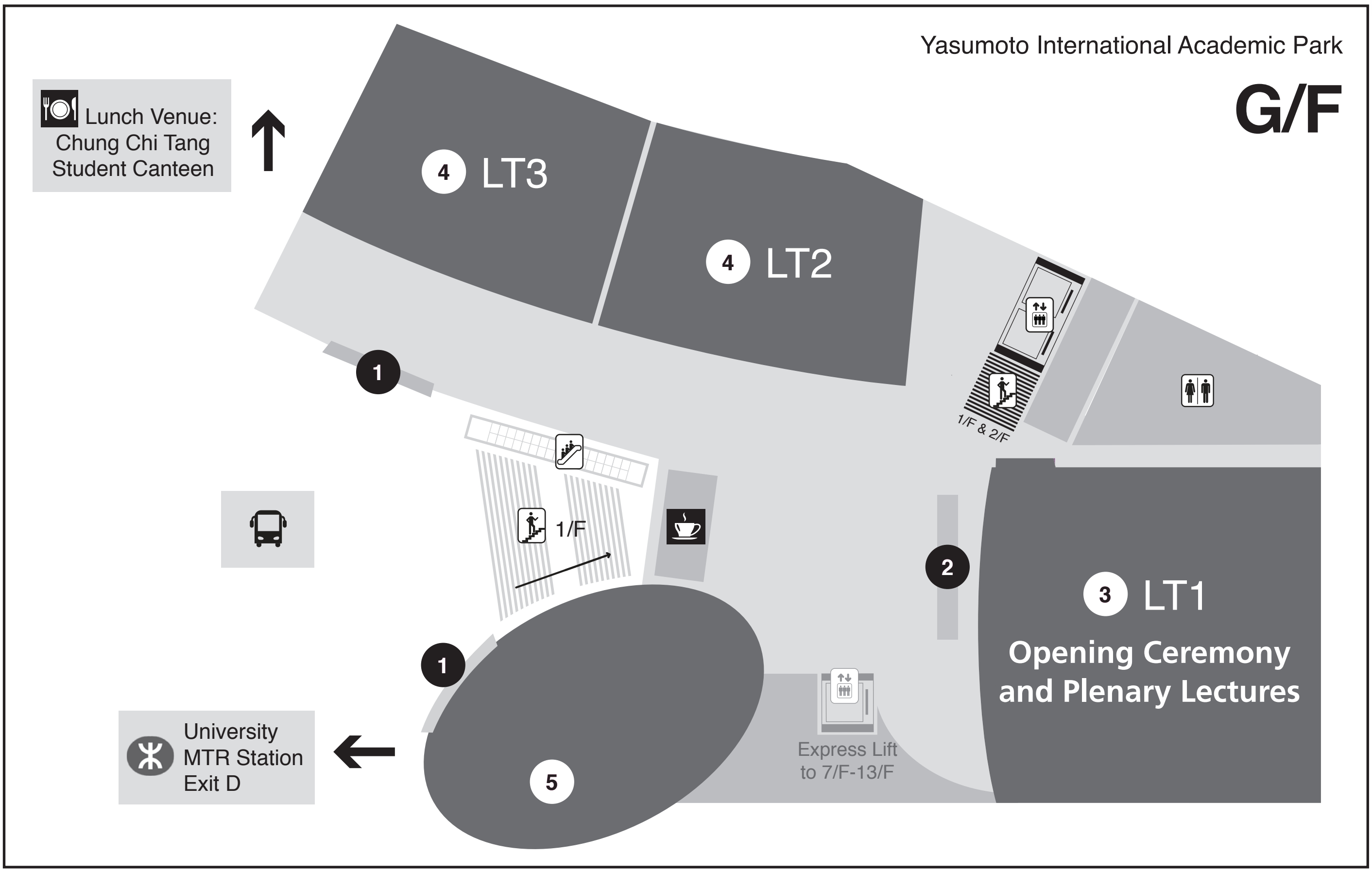
**Mong Kok station and Mong Kok East station are DIFFERENT.*

CUHK Campus Map



CUHK Campus Map
www.cuhk.edu.hk/english/campus/cuhk-campus-map.html

Floor Maps



1 Entrances

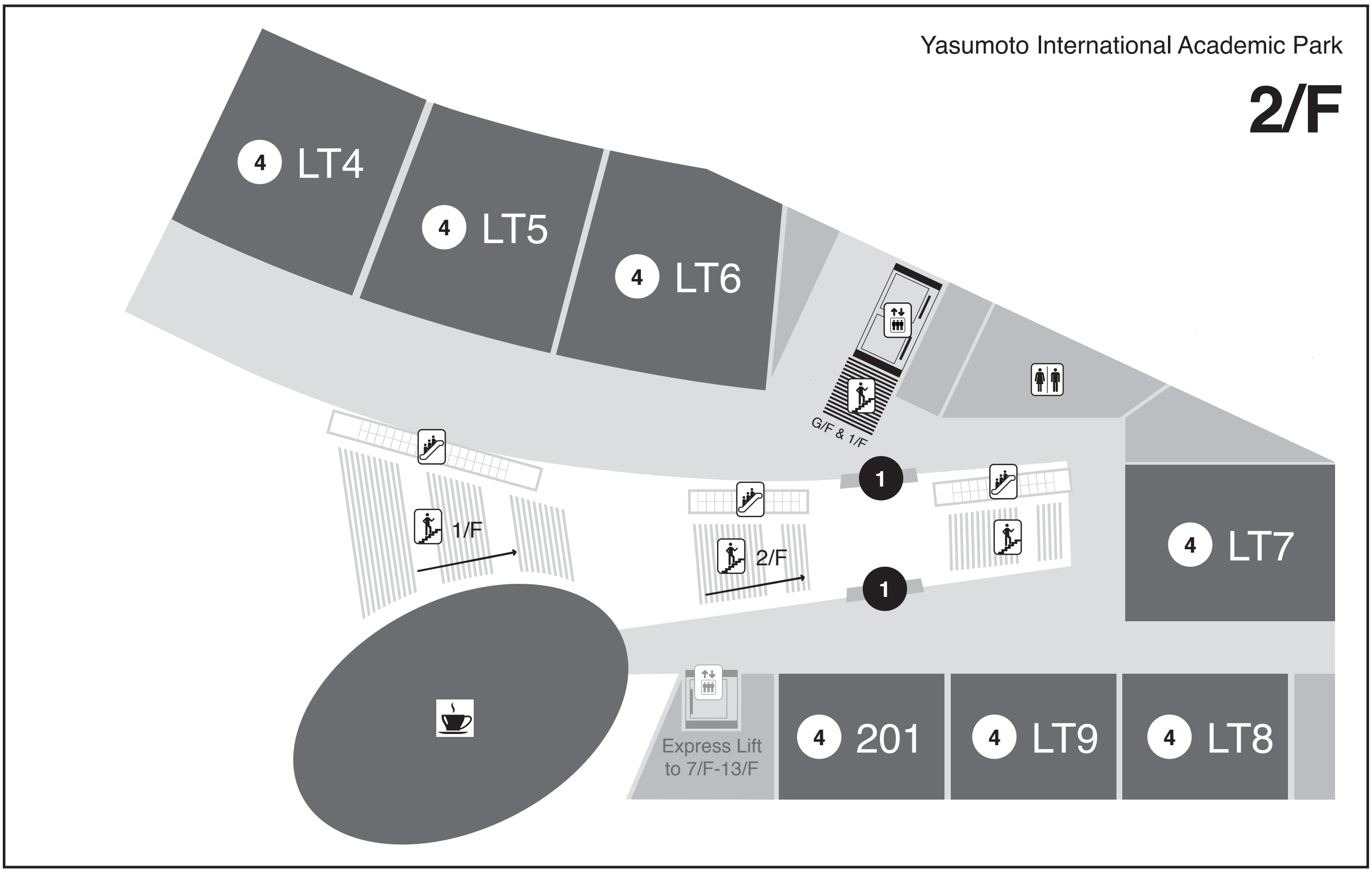
2 Registration and Enquiry Counter

3 **Opening Ceremony and Plenary Lectures**

- Lecture Theatre 1, G/F

4 **Oral Presentations: Distinguished Lectures, Invited Paper Sessions, Topic Contributed Paper Sessions and Contributed Paper Sessions**

- Lecture Theatre 2, G/F
- Lecture Theatre 3, G/F
- Lecture Theatre 4, 2/F
- Lecture Theatre 5, 2/F
- Lecture Theatre 6, 2/F
- Lecture Theatre 7, 2/F
- Lecture Theatre 8, 2/F
- Lecture Theatre 9, 2/F
- Room 201, 2/F
- Room 403, 4/F
- Room 405, 4/F
- Room 505, 5/F



❺ **Poster Presentations**

🛗 Lifts, Escalators, Staircases

🛗 Express Lift to 7/F-13/F

🚻 Toilets

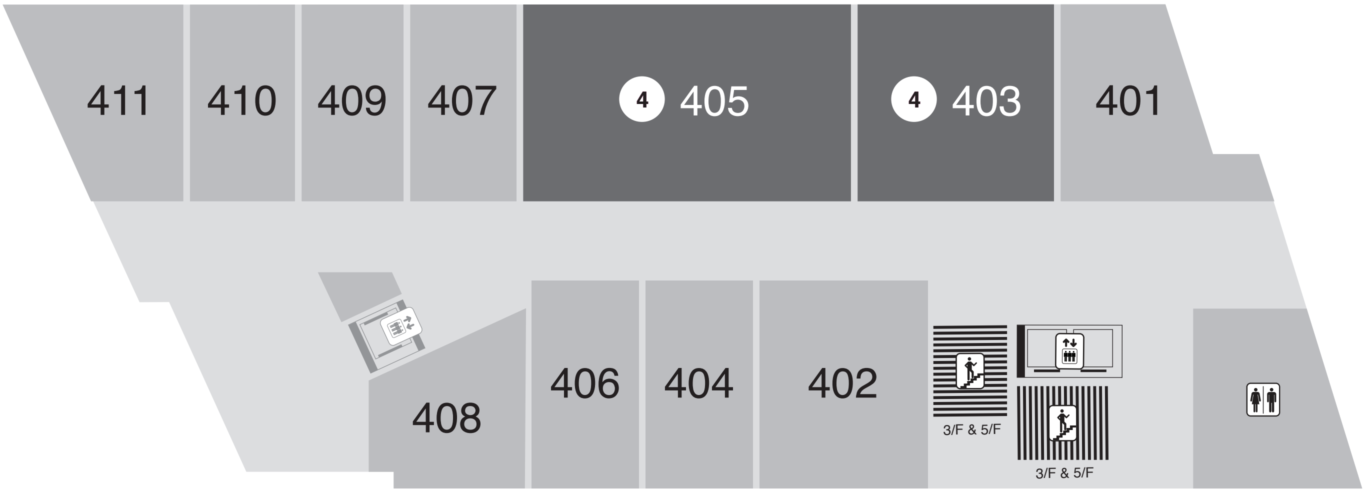
☕ Coffee Breaks

🍽 Lunch Venue

🚌 Conference Dinner and Excursion Buses Boarding Point

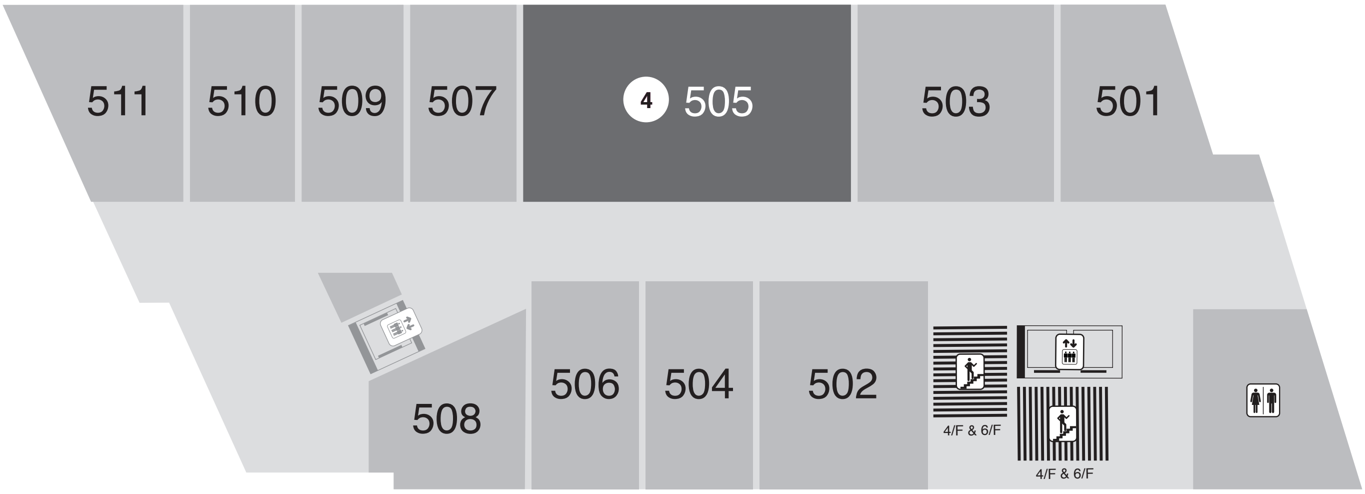
Yasumoto International Academic Park

4/F



Yasumoto International Academic Park

5/F



Program

Day 1 Mon, June 27

■ **08:00-17:00** | Registration and Enquiry — YIA G/F

■ **09:00-09:40** | **Opening Ceremony** — YIA G/F LT1

■ **09:40-10:00** | Coffee Break — YIA G/F & 2/F

■ **10:00-11:00** | **Plenary Lecture 1** — YIA G/F LT1
Understanding Importance Sampling
Plenary Speaker: Persi Diaconis (Stanford University)
Chair: Qi-Man Shao (The Chinese University of Hong Kong)

■ **11:10-12:10** | **Plenary Lecture 2** — YIA G/F LT1
Low Rank Structure in Highly Multivariate Models
Plenary Speaker: Iain Johnstone (Stanford University)
Chair: Ming-Yen Cheng (National Taiwan University)

■ **12:10-13:30** | Lunch — Chung Chi Tang Student Canteen

■ **13:30-15:10**

Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics

DL07 (YIA G/F LT2)

Sponsor: IISA
Chair: Amarjit Budhiraja (The University of North Carolina at Chapel Hill)

Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics
Distinguished Lecturer: Rabi Bhattacharya (The University of Arizona)

Invited papers:

1. **Dimension Reduction on Tori and Polyspheres**
   Stephan Huckemann (University Göttingen)
2. **Nonparametric Regression on Manifolds**
   Lizhen Lin (The University of Texas at Austin)

Recent Advances in Covariance Estimation

DL09 (YIA G/F LT3)

Sponsor: Korea
Chair: Woncheol Jang (Seoul National University)

Covariance Estimation with the Positive Definite Constraint
Distinguished Lecturer: Ja-Yong Koo (Korea University)

Invited papers:

1. **Local Distance Regression Model for Manifold-valued Data**
   Hongtu Zhu (The University of North Carolina at Chapel Hill)
2. **Solution Path of Condition-number Regularization**
   Joong-Ho Won (Seoul National University)

Recent Advances on Analysis of High-Dimensional Data

IP01 (YIA 2/F LT4)

Sponsor: IMS
Organizer: Runze Li (The Pennsylvania State University)
Chair: Runze Li (The Pennsylvania State University)

Invited papers:

1. **De-biasing Regularized Estimators With High-dimensional Data**
   Cun-Hui Zhang (Rutgers University)
2. **CoCoLasso for High-dimensional Error-in-variables Regression**
   Hui Zou (University of Minnesota)
3. **Slow Kill for Big Data Screening**
   Yiyuan She (Florida State University)
4. **Nonparametric Two-sample Test in Ultra-high Dimension**
   Lan Wang (University of Minnesota)

Robust Statistical Methods

IP34 (YIA 2/F LT5)

Sponsor: India
Organizer: Subhra Sankar Dhar (Indian Institute of Technology, Kanpur)
Chair: Subhra Sankar Dhar (Indian Institute of Technology, Kanpur)

Invited papers:

1. **Big Data World: Wide Consensus in Estimation using Parallelized Inference**
   Juan Antonio Cuesta-Albertos (Universidad de Cantabria)
2. **On Bootstrap and Robustness of Regularized Kernel Based Methods**
   Andreas Christmann (University of Bayreuth)
3. **Robust Estimation of Precision Matrices under Cellwise Contamination**
   Garth Tarr (The University of Newcastle)
4. **Penalized Weighted Least Squares for Outlier Detection and Robust Regression**
   Xiaoli Gao (The University of North Carolina at Greensboro)

Multivariate Smoothing

IP37 (YIA 2/F LT6)

Sponsor: Australia and New Zealand
Organizer: Martin Hazelton (Massey University)
Chair: Spiridon Ivanov Penev (The University of New South Wales)

Invited papers:

1. **Unconstrained Multivariate Bandwidth Selection for Density and Density Derivative Estimation**
   Jose E. Chacon (University of Extremadura)
2. **Asymptotics of Variable-bandwidth Kernel Density and Density-ratio Estimation for Planar Point Patterns in Epidemiology**
   Tilman M. Davies (University of Otago)
3. **Probit Transformation for Nonparametric Kernel Estimation of the Copula Density**
   Gery Geenens (The University of New South Wales)
4. **Multivariate Log-density Estimation with Applications to Approximate Likelihood Inference**
   Martin Hazelton (Massey University)

Complex Data Analysis

IP21 (YIA 2/F LT7)

Sponsor: Chinese Society of Probability and Statistics
Organizer: Niansheng Tang (Yunnan University)
Chair: Liping Zhu (Renmin University of China)

Invited papers:

1. **Functional Single-index Model for Functional Data**
   Zhongyi Zhu (Fudan University)
2. **On Improving Efficiency by Borrowing Information across Quantiles**
   Yuanyuan Lin (The Chinese University of Hong Kong)
3. **Nonparametric Model for Panel Data with Fixed Effects and Locally Stationary Regressors**
   Tao Huang (Shanghai University of Finance and Economics)

Complex Data Analysis: Theory and Methods

IP51 (YIA 2/F LT8)

Organizer: Yang Feng (Columbia University)
Chair: Yang Feng (Columbia University)

Invited papers:

1. **Screening and Feature Selection in High-dimensional Fisher Discriminant Analysis**
   Ming-Yen Cheng (National Taiwan University)
2. **Adaptive Sparse Non-linear Metric Learning via Boosting**
   Tian Zheng (Columbia University)
3. **QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization**
   Lucy Xia (Stanford University)
4. **Estimating Network Edge Probabilities by Neighborhood Smoothing**
   Ji Zhu (University of Michigan)

Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques II: Robust SEM, Multilevel SEM, Meta-Analytic SEM, Latent Profile Analysis, and Cross-Classified SEM

TCP15 (YIA 2/F LT9)

Organizer: Oi-Man Kwok (Texas A&M University)
Chair: Oi-Man Kwok (Texas A&M University)

Topic contributed papers:

1. **Meta-analysis: A Structural Equation Modeling Approach**
   Mike W.-L. Cheung (National University of Singapore)
2. **The Impact of ICC on the Effectiveness of Level-specific Fit Indices in Multilevel Structural Equation Modeling: A Monte Carlo Study**
   Hsien-Yuan Hsu (University of Mississippi)
3. **Using Multivariate-t-based Maximum Likelihood for Robust Structural Equation Modeling**
   Hok Chio Lai (University of Cincinnati)
4. **Applying Latent Profile Analysis on the Development of Self-control and Self-esteem Configuration Among Adolescents**
   Yuan-Hsuan Lee (National Taichung University of Education)
5. **Testing Mediation Effects in Cross-classified Multilevel Data**
   Wen Luo (Texas A&M University)

Recent Advances and Challenges in Analysis of Complex Biomedical Data

TCP16 (YIA 2/F 201)

Organizer: Sijian Wang (University of Wisconsin-Madison)
Chair: Lingsong Zhang (Purdue University)

Topic contributed papers:

1. **Latent Class Modeling using Matrix-valued Covariates with Application to Identifying Early Placebo Responders Based on EEG Signals**
   Bei Jiang (University of Alberta)
2. **Quantile Regression with Varying Coefficients for Functional Responses**
   Linglong Kong (University of Alberta)
3. **A Multi-scale Spatial Point Process Model for Stroke Lesion Segmentation on Multimodal MRI Data**
   Huiyan Sang (Texas A&M University)
4. **Integrative Analysis of High-dimensional Genomic Data**
   Sijian Wang (University of Wisconsin-Madison)

Recent Advances in Object Data Analysis

TCP17 (YIA 4/F 403)

Organizer: Jie Peng (University of California)
Chair: Sungkyu Jung (University of Pittsburgh)

Topic contributed papers:

1. **Sequential Change-point Detection Based on Nearest Neighbors**
   Hao Chen (University of California)
2. **Small Circle Distributions for Estimation of Rotational Axis from Directional Data**
   ByungWon Kim (University of Pittsburgh)
3. **Statistical Analysis of Trajectories on Riemannian Manifolds**
   Jingyong Su (Texas Tech University)
4. **Multiscale Modeling of Hi-C Data**
   Rachel Wang (Stanford University)

Machine Learning Session 1

CP01 (YIA 4/F 405)

Chair: Xiaodan Fan (The Chinese University of Hong Kong)

Contributed papers:

1. **Bayesian Bandwidth Selection in Nonparametric Models Based on Local Polynomial Regression**
   Zhongcheng Han (Southeast University)
2. **Clustering Functional Data using Projection**
   Tung Pham (The University of Melbourne)
3. **Optimal Estimation of Derivatives in Nonparametric Regression**
   Wenlin Dai (King Abdullah University of Science and Technology)
4. **Modeling and Forecasting On-line Auction Prices: A Semi-parametric Regression Analysis**
   Weiwei Liu (Lanzhou University)
5. **Sparse Regularization for Multiclass Functional Logistic Regression**
   Hidetoshi Matsui (Shiga University)

■ 15:10-15:30 | Coffee Break

YIA G/F & 2/F

■ 15:30-17:10

Statistical Analysis for Social Network Data

DL10 (YIA G/F LT2)

Sponsor: Chinese Society of Probability and Statistics
Chair: Hoi Ying Wong (The Chinese University of Hong Kong)

Network Vector Autoregression
Distinguished Lecturer: Hansheng Wang (Peking University)

Invited papers:

1. **Least Squares Estimation of Spatial Autoregressive Models for Large-scale Social Networks**
   Danyang Huang (Renmin University of China)
2. **Multivariate Spatial Autoregression for Large Scale Social Network**
   Xuening Zhu (Peking University)

Sparse Learning from Covariance

DL17 (YIA G/F LT3)

Sponsor: IMS
Chair: Cun-Hui Zhang (Rutgers University)

Robust Statistical Learning from Covariance
Distinguished Lecturer: Jianqing Fan (Princeton University)

Invited papers:

1. **Discussion of the Distinguished Lecture**
   Harrison Zhou (Yale University)
2. **Large-scale Mean-variance Portfolio Optimization**
   Xinghua Zheng (The Hong Kong University of Science and Technology)

Big Data Integration

IP07 (YIA 2/F LT4)

Sponsor: IMS
Organizer: Jason Fine (The University of North Carolina at Chapel Hill)
Chair: Jason Fine (The University of North Carolina at Chapel Hill)

Invited papers:

1. **Integrative Statistical Analysis and Exploration of Mixed-type Massive Data**
   Susan Wilson (The University of New South Wales and The Australian National University)
2. **Shared Informative Factor Models for Integration of Multi-platform Bioinformatic Data**
   Jianhua Hu (The University of Texas MD Anderson Cancer Center)
3. **Fusion Learning in Data Integration**
   Peter XK Song (University of Michigan)
4. **Optimal Estimation for Quantile Regression with Functional Response**
   Xiao Wang (Purdue University)

Longitudinal Data Analysis

IP12 (YIA 2/F LT5)

Sponsor: ICSA
Organizer: Jialiang Li (National University of Singapore)
Chair: Jialiang Li (National University of Singapore)

Invited papers:

1. **On Last Observation Carried Forward and Asynchronous Longitudinal Regression Analysis**
   Hongyuan Cao (University of Missouri)
2. **Variable Selection for Fixed and Random Effects in Generalized Linear Mixed Models**
   Liming Xiang (Nanyang Technological University)
3. **Efficient Estimation in Semivarying Coefficient Models for Longitudinal / Clustered Data**
   Toshio Honda (Hitotsubashi University)
4. **Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories with Application to Cocaine Abuse Treatment Data**
   Yehua Li (Iowa State University)

Limit Theorems in Stochastic Analysis

IP15 (YIA 2/F LT6)

Sponsor: IISA
Organizer: Arnab Ganguly (Louisiana State University)
Chair: Amarjit Budhiraja (The University of North Carolina at Chapel Hill)

Invited papers:

1. **Moderate Deviation Principles for Stochastic Differential Equations**
   Arnab Ganguly (Louisiana State University)
2. **On a Rescaling Transformation for Stochastic Partial Differential Equations**
   Michael Roeckner (Bielefeld University)
3. **Large Deviations for Multi-scale Jump-diffusions**
   Rohini Kumar (Wayne State University)
4. **Dissipation and High Disorder**
   Kunwoo Kim (Pohang University of Science and Technology)

Nonparametric Inference in Econometric Models

IP20 (YIA 2/F LT7)

Sponsor: Korea
Organizer: Kyusang Yu (Konkuk University)
Chair: Kyusang Yu (Konkuk University)

Invited papers:

1. **Operational Time and In-sample Density Forecasting**
   Young K. Lee (Kangwon National University)
2. **A New Estimator for Stochastic Frontier Models**
   Hohsuk Noh (Sookmyung Women's University)
3. **Partial Identification in Regression Discontinuity Designs with Manipulated Running Variables**
   Christoph Rothe (Columbia University)
4. **A Semiparametric Intraday GARCH-X Model**
   Melanie Schienle (Karlsruhe Institute of Technology)

Recent Advances and Challenges of Big Data Inference with Complex Structures

IP57 (YIA 2/F LT8)

Organizer: Dan Yang (Rutgers University)
Chair: Dan Yang (Rutgers University)

Invited papers:

1. **Linear Regression Analysis in a Model Free Setting**
   Linda Zhao (University of Pennsylvania)
2. **The Power Prior: Theory and Applications**
   Ming-Hui Chen (University of Connecticut)
3. **Two-sample Tests for High-dimensional Linear Regression with an Application to Gene and Environment Interactions**
   Yin Xia (The University of North Carolina at Chapel Hill)
4. **Nested Nonnegative Cone Analysis**
   Lingsong Zhang (Purdue University)

Analysis of Reliability and Survival Data

TCP12 (YIA 2/F LT9)

Organizer: Tony Hon Keung Ng (Southern Methodist University)
Chair: Tony Hon Keung Ng (Southern Methodist University)

Topic contributed papers:

1. **Lifetime Inference for Highly Reliable Products Based on Skew-normal Accelerated Destructive Degradation Test Model**
   Chien-Tai Lin (Tamkang University)
2. **Proportional Hazards Model for Accelerated Life Testing Data from One-shot Devices**
   Man Ho Ling (The Hong Kong Institute of Education)
3. **Inference for the Generalized Pareto Distribution and its Application**
   Hideki Nagatsuka (Chuo University)
4. **The EM Algorithm for One-shot Device Testing with Competing Risk under Different Lifetimes Distributions**
   Hon Yiu So (McMaster University)
5. **Strategic Allocation of Test Units in an Accelerated Degradation Test Plan**
   Zhisheng Ye (National University of Singapore)

Theory and Applications of Tensor Variate Data Analysis

TCP13 (YIA 2/F 201)

Organizer: Toshio Sakata (Kyushu University)
Chair: Kohei Adachi (Osaka University)

Topic contributed papers:

1. **Inference for Tensor Elliptical Distributions**
   Mohammad Arashi (Shahrood University of Technology)
2. **Universal Subspaces for Local Unitary Groups of Fermionic Systems**
   Lin Chen (Beihang University)
3. **Sparse Three-way PCA for Selecting the Optimal Model Between Tucker2 and Parafac**
   Hiroki Ikemoto (Osaka University)
4. **One-sided Tests for Tensor Variate Normal Distributions**
   Manabu Iwasa (Kumamoto University)
5. **Typical Ranks of Tensors Over the Real Number Field and Determinantal Ideals**
   Mitsuhiro Miyazaki (Kyoto University of Education)

Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques I: Measurement Invariance and Latent Growth Models

TCP14 (YIA 4/F 403)

Organizer: Oi-Man Kwok (Texas A&M University)
Chair: Oi-Man Kwok (Texas A&M University)

Topic contributed papers:

1. **Model Specification Search for Identifying the Optimal Growth Trajectory in Latent Growth Models**
   MinJung Kim (The University of Alabama)
2. **Multilevel Factorial Invariance in Ordered Categorical Measures**
   Ehri Ryu (Boston College)
3. **Testing Longitudinal Measurement Invariance using Majority Votes through a Sequential Procedure**
   Jiun-Yu Wu (National Chiao Tung University)
4. **Testing Factorial Invariance with Severely Unbalanced Samples**
   Myeongsun Yoon (Texas A&M University)

Machine Learning Session 2

CP02 (YIA 4/F 405)

Chair: Tung Pham (The University of Melbourne)

Contributed papers:

1. **Comparing Two Nonparametric Regression Curves in the Presence of Long Memory in Covariates and Errors**
   Fang Li (Indiana University-Purdue University Indianapolis)
2. **Robust Regression for Highly Corrupted Response by Shifting Outliers**
   Yoonsuh Jung (The University of Waikato)
3. **Schwarz-type Model Comparison for LAQ Models**
   Shoichi Eguchi (Kyushu University)
4. **Kernel Entropy Estimation for Linear Processes**
   Hailin Sang (The University of Mississippi)

Day 2 Tue, June 28

■ **08:00-17:00** | Registration and Enquiry YIA G/F

■ **08:30-10:10**

Recent Development of Statistical Inference Under Dimension Reduction Structure

DL13 (YIA G/F LT2)

Sponsor: Local Host
Chair: Xin Yuan Song (The Chinese University of Hong Kong)

Adaptive-to-model Test for Parametric Single-index Models: A Dimension Reduction Approach
Distinguished Lecturer: Lixing Zhu (Hong Kong Baptist University)

Invited papers:

1. **Sufficient Dimension Reduction for Longitudinal Data**
   Annie Peiyong Qu (University of Illinois at Urbana-Champaign)
2. **A Post-screening Diagnostic Study in Sufficient Dimension Reduction for Ultrahigh Dimensional Data**
   Liping Zhu (Renmin University of China)

Special Session in Memory of Peter Hall

IP60 (YIA G/F LT3)

Sponsor: IMS
Organizer: Alan Welsh (Australian National University)
Chair: Alan Welsh (Australian National University)

Invited Speakers:

- Susan Wilson (The University of New South Wales and The Australian National University)
- Matt Wand (University of Technology Sydney)
- Jianqing Fan (Princeton University)
- Byeong Park (Seoul National University)
- Bingyi Jing (The Hong Kong University of Science and Technology)
- Ming-Yen Cheng (National Taiwan University)

Models and Inference for Complex Extreme Events IP16 (YIA 2/F LT4)

Sponsor: Bernoulli Society
Organizer: Anthony Davison (Ecole Polytechnique Fédérale de Lausanne)
Chair: Richard A. Davis (Columbia University)

Invited papers:

1. **Generalized Additive Modeling of Nonstationary Multivariate Extremes**
   Miguel de Carvalho (Pontificia Universidad Católica de Chile)
2. **Exact Simulation of Max-stable Processes**
   Sebastian Engelke (Ecole Polytechnique Fédérale de Lausanne)
3. **Full Likelihood Inference For Max-stable Distributions Based on a Stochastic EM Algorithm**
   Raphael Huser (King Abdullah University of Science and Technology)
4. **Probabilities of Concurrent Extremes**
   Stilian Stoev (University of Michigan)

Asymptotic Theory IP28 (YIA 2/F LT5)

Sponsor: Local Host
Organizer: Qi-Man Shao (The Chinese University of Hong Kong)
Chair: Qi-Man Shao (The Chinese University of Hong Kong)

Invited papers:

1. **Parisi Formula, Disorder Chaos and Fluctuation for the Ground State Energy in the Spherical Mixed p-spin Models**
   Wei-Kuo Chen (University of Minnesota)
2. **Testing Independence with High-dimensional Correlated Samples**
   Weidong Liu (Shanghai Jiao Tong University)
3. **Berry-Esseen Bound for Exchangeable Pairs**
   Zhuosong Zhang (The Chinese University of Hong Kong)
4. **Spurious Discoveries in High Dimension**
   Wenxin Zhou (Princeton University)

Distributions on Convex Cones and Sets of Permutations IP32 (YIA 2/F LT6)

Sponsor: Japan
Organizer: Satoshi Kuriki (The Institute of Statistical Mathematics)
Chair: Donald Richards (The Pennsylvania State University)

Invited papers:

1. **Multivariate Distribution Theory by Holonomic Gradient Method**
   Akimichi Takemura (Shiga University)
2. **Generation of Random Permutations: Algorithms, Implementation and Probabilistic Analysis**
   Hsien-Kuei Hwang (Academia Sinica)
3. **Some Distributions Associated with the Cone of Positive Semidefinite Matrices and their Applications**
   Satoshi Kuriki (The Institute of Statistical Mathematics)
4. **Totally Positive Exponential Families**
   Caroline Uhler (Massachusetts Institute of Technology)

Survey Sampling and Related Methodologies

IP35 (YIA 2/F LT7)

Sponsor: India
Organizer: Sanjay Chaudhuri (National University of Singapore)
Chair: Malay Ghosh (University of Florida)

Invited papers:

1. **An Empirical Likelihood Based Estimator for Respondent Driven Sampled Data**
   Sanjay Chaudhuri (National University of Singapore)
2. **Small Area Model Selection**
   Singdhansu Chatterjee (University of Minnesota)
3. **Multiple Imputation and/or Calibration in Two-phase Designs?**
   Thomas Lumley (The University of Auckland)
4. **Multiple Imputation using the Weighted Finite Population Bayesian Bootstrap**
   Michael Elliott (University of Michigan)

Statistics and Computing for Complex Dependent Systems

TCP03 (YIA 2/F LT8)

Organizer: Kengo Kamatani (Osaka University)
Chair: Teppei Ogihara (The Institute of Statistical Mathematics)

Topic contributed papers:

1. **Robust Estimation for Sparse Gaussian Graphical Modeling**
   Kei Hirose (Osaka University)
2. **Multilevel Sequential Monte Carlo Samplers**
   Ajay Jasra (National University of Singapore)
3. **On Asymptotics of Multivariate Non-Gaussian Quasi-likelihood**
   Hiroki Masuda (Kyushu University)
4. **Some Properties of the Mixed Preconditioned Crank-Nicolson Algorithm**
   Kengo Kamatani (Osaka University)

Stochastic Processes and Related Topics

TCP08 (YIA 2/F LT9)

Organizer: Dongsheng Wu (The University of Alabama in Huntsville)
Chair: Dongsheng Wu (The University of Alabama in Huntsville)

Topic contributed papers:

1. **Small Value Probabilities for Supercritical Multitype Branching Processes**
   Weijuan Chu (Nanjing University)
2. **Some Aspects of the Rosenblatt Sheet**
   Guangjun Shen (Anhui Normal University)
3. **Stepwise Estimation of Ergodic Levy Driven Stochastic Differential Equation**
   Yuma Uehara (Kyushu University)
4. **Quantile Regression Process: Nonparametric and Partially Linear Asymptotics**
   Shih-Kang Chao (Purdue University)

Recent Advances in Time-to-Event Analysis

TCP23 (YIA 2/F 201)

Organizer: Tony Sit (The Chinese University of Hong Kong)
Chair: Chi Wing George Chu (The Chinese University of Hong Kong)

Topic contributed papers:

1. **Pseudo Value Method for Ultra High-dimensional Semiparametric Models with Life-time Data**
   Tony Sit (The Chinese University of Hong Kong)
2. **Efficient Estimation and Inference of Quantile Regression Under Biased Sampling**
   Gongjun Xu (University of Minnesota)
3. **Confidence Intervals for High-dimensional Cox Models**
   Yi Yu (University of Cambridge)
4. **End-point Sampling**
   Wen Yu (Fudan University)

Geometric Aspects of Statistical Methods

TCP06 (YIA 4/F 403)

Organizer: Tomonari Sei (The University of Tokyo)
Chair: Tomonari Sei (The University of Tokyo)

Topic contributed papers:

1. **A Decision Theoretic Property of Conditional Normalized Maximum Likelihood Distribution**
   Yoshihiro Hirose (The University of Tokyo)
2. **Information Geometry of Anomalous Statistics**
   Hiroshi Matsuzoe (Nagoya Institute of Technology)
3. **Symmetries in Experimental Design and Linear Estimators**
   Kentaro Tanaka (Seikei University)
4. **On Submanifolds of Textile Set**
   Ushio Tanaka (Osaka Prefecture University)

Data Order Structure Session 1

CP14 (YIA 4/F 405)

Chair: Chun Yip Yau (The Chinese University of Hong Kong)

Contributed papers:

1. **Quasi Hidden Markov Model and its Applications in Change-point Problems**
   Zhengxiao Wu (Singapore Management University)
2. **Autoregressive Models using Geometric Stable Distributions**
   Kuttykrishnan Adavalath Puthiyaveetil (Sir Syed College)
3. **Bayesian Local Influence Analysis for Generalized Autoregressive Conditional Heteroscedasticity Model with Empirical Applications**
   Hongxia Hao (Southeast University)
4. **On Statistical Tests for Change Points of Poisson Processes**
   Christian Farinetto (Université du Maine)

Design of Experiments

DL11 (YIA 5/F 505)

Sponsor: Chinese Statistical Association (Taiwan)
Chair: Tsung-Chi Cheng (National Chengchi University)

Optimal Design of fMRI Experiments
Distinguished Lecturer: Ching-Shui Cheng (Academia Sinica)

Invited papers:

1. **Experimental Designs for Functional MRI with Uncertain Model Matrix**
   Ming-Hung Kao (Arizona State University)
2. **Optimal Design of fMRI Experiments Using Circulant (Almost-)Orthogonal Arrays**
   Frederick Kin Hing Phoa (Academia Sinica)

■ 10:10-10:30 | Coffee Break

YIA G/F & 2/F

■ 10:30-12:10

Building Bridges: New Bayesian Insights into Old Problems

DL02 (YIA G/F LT2)

Sponsor: IMS
Chair: Ming-Hui Chen (University of Connecticut)

Building Bridges: New Bayesian Insights into Old Problems Reflections on Bayesian Priors
Distinguished Lecturer: Kerrie Mengersen (Queensland University of Technology)

Invited papers:

1. **Simultaneous Estimation of Spatial Frequency Fields and Measurement Locations, with an Application to Spatial Location of Late Middle English Texts**
   Geoff Nicholls (University of Oxford)
2. **Praising the Prior**
   Eric-Jan Wagenmakers (University of Amsterdam)

Modern Developments in Multivariate Data

IP02 (YIA G/F LT3)

Sponsor: IMS
Organizer: Debajyoti Sinha (Florida State University)
Chair: Yiyuan She (Florida State University)

Invited papers:

1. **Skew-symmetric Models for Highly Skewed Clustered Data**
   Debajyoti Sinha (Florida State University)
2. **Bayes Theory and Methods for Large Networks**
   Debdeep Pati (Florida State University)
3. **Global-Local Shrinkage Priors for Variable Selection and Estimation**
   Malay Ghosh (University of Florida)
4. **Parsimonious Tensor Response Regression with Applications to Neuroimaging Analysis**
   Xin Zhang (Florida State University)

Statistical Analysis of Complex Correlated Data

IP03 (YIA 2/F LT4)

Sponsor: IMS
Organizer: Peter XK Song (University of Michigan)
Chair: Peter XK Song (University of Michigan)

Invited papers:

1. **Assessing Genomic Risk for Learning Problems with Neuroimaging Data**
   Heping Zhang (Yale University School of Public Health)
2. **Simultaneous Feature Selection and Precision Matrix Estimation in High-dimensional Multivariate Regression Models**
   Zehua Chen (National University of Singapore)
3. **Residual-based Model Diagnosis Methods for Mixture Cure Models**
   Yingwei Peng (Queen's University)
4. **Composite Quantile Regression for Correlated Data**
   Heng Lian (The University of New South Wales)

Stein's Method and Random Graphs

IP44 (YIA 2/F LT5)

Organizer: Adrian Roellin (National University of Singapore)
Chair: Adrian Roellin (National University of Singapore)

Invited papers:

1. **Change Point Detection in Evolving Network Models**
   Shankar Bhamidi (The University of North Carolina at Chapel Hill)
2. **Rates of Convergence for Multivariate Normal Approximation with Applications to Dense Graphs**
   Xiao Fang (National University of Singapore)
3. **Limit Behavior of Some Polya Urn Models Associated to Preferential Attachment Graphs and Random Trees**
   Nathan Ross (The University of Melbourne)
4. **Bounds on the Condensation Threshold in Stochastic Block Models**
   Joe Neeman (University of Texas at Austin / University of Bonn)

New Frontiers of Longitudinal/Clustered Data Analysis

IP18 (YIA 2/F LT6)

Sponsor: Korea
Organizer: Mi-Ok Kim (Cincinnati Children's Hospital Medical Center)
Chair: Yehua Li (Iowa State University)

Invited papers:

1. **Disease Progress Monitoring Based on Bayesian Joint Modeling of Left-truncated Longitudinal and Survival Outcomes**
   Mi-Ok Kim (Cincinnati Children's Hospital Medical Center)
2. **A Shared Parameter Model for Curve Registration in the Presence of Informative Dropout**
   Sarah J. Ratcliffe (University of Pennsylvania)
3. **Marginal Zero-inflated Regression Models for Cross-sectional and Clustered Count Data**
   Daniel Hall (University of Georgia)
4. **Doubly Robust Generalized Estimating Equations with Consistent Variance Estimator when One Auxiliary Model is Misspecified**
   Soeun Kim (The University of Texas Health Science Center at Houston)

Dimension Reduction and Variable Selection for High-Dimensional Data

IP38 (YIA 2/F LT7)

Sponsor: Australia and New Zealand
Organizer: Inge Koch (The University of Adelaide)
Chair: Valentin Patilea (ENSAI)

Invited papers:

1. **Pseudo Sufficient Dimension Reduction and Variable Selection**
   Xiangrong Yin (University of Kentucky)
2. **On the Effective Number of Principal Components in HDLSS Context**
   Sungkyu Jung (University of Pittsburgh)
3. **Kernel Naive Bayes for High Dimensional Pattern Recognition**
   Kanta Naito (Shimane University)
4. **Classification and Variable Selection for High-dimensional Data with Application to Proteomics**
   Inge Koch (The University of Adelaide)

Recent Advances in Analysis of Complex Data

IP54 (YIA 2/F LT8)

Organizer: Xin Yuan Song (The Chinese University of Hong Kong)
Chair: Xin Yuan Song (The Chinese University of Hong Kong)

Invited papers:

1. **Bayesian Neural Networks for Personalized Medicine**
   Faming Liang (University of Florida)
2. **A Concave Pairwise Fusion Approach to Subgroup Analysis**
   Jian Huang (The University of Iowa)
3. **Bayesian Shape-restricted Analysis of Complex Data using Gaussian Process Priors**
   Taeryon Choi (Korea University)
4. **Ball Divergence: Nonparametric Two Sample Test**
   Xueqin Wang (Sun Yat-sen University)

Recent Advances in Statistical Genetics

TCP18 (YIA 2/F LT9)

Organizer: Hong Zhang (Fudan University)
Chair: Zhaohai Li (The George Washington University)

Topic contributed papers:

1. **A Principal Score Test for Analyzing Multiple Genetic Markers**
   Jinbo Chen (University of Pennsylvania)
2. **Extension of the Peters-Belson Method to Estimate Health Disparities Among Multiple Groups using Logistic Regression with Survey Data**
   Yan Li (University of Maryland)
3. **A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations**
   Jianxin Shi (National Cancer Institute)
4. **Strategies for Conducting Multi-locus Analysis using Single-locus Summary Statistics**
   Kai Yu (National Cancer Institute)

Recent Developments in Model Checking

TCP19 (YIA 2/F 201)

Organizer: Xu Guo (Nanjing University of Aeronautics and Astronautics)
Chair: Maolin Pan (Nanjing University)

Topic contributed papers:

1. **Enhancements of Nonparametric Generalized Likelihood Ratio Test: Bias-correction and Dimension Reduction**
   Xu Guo (Nanjing University of Aeronautics and Astronautics)
2. **A Robust Adaptive-to-model Enhancement Test for Parametric Single-index Models**
   Cui Zhen Niu (Renmin University of China)
3. **Dimension Reduction-based Model Checking for Survival Models**
   Jingke Zhou (Ningbo University)
4. **An Adaptive-to-model Test for Partially Parametric Single-index Models**
   Xuehu Zhu (Xi'an Jiaotong University)

Bayesian Modeling with Complex Survey Data

TCP10 (YIA 4/F 403)

Organizer: Cici Chen Bauer (Brown University)
Chair: Tony Sit (The Chinese University of Hong Kong)

Topic contributed papers:

1. **Robust Bayesian Models for Surveys with Missing Data and External Information**
   Sahar Zangeneh (Fred Hutchinson Cancer Research Center)
2. **Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question**
   Guo-Liang Tian (The University of Hong Kong)
3. **Bayesian Empirical Likelihood Methods for Complex Surveys**
   Changbao Wu (University of Waterloo)
4. **Bayesian Spatial Hierarchical Models for Small Estimation with Complex Survey Designs**
   Cici Chen Bauer (Brown University)

Machine Learning Session 3

CP03 (YIA 4/F 405)

Chair: Yoonkyung Lee (The Ohio State University)

Contributed papers:

1. **On a General Procedure for Constructing Confidence Sets under Partially Identified Models**
   Han Jiang (The University of Hong Kong)
2. **Robust Non-convex Penalized Linear Regression with Algorithmic and Statistical Convergence**
   Shota Katayama (Tokyo Institute of Technology)
3. **Adaptive Kernel-based FPCA for Functional Generalized Linear Models**
   Guangbao Guo (Shandong University)

■ **12:10-13:30** | Lunch — Chung Chi Tang Student Canteen

■ **13:00-13:45** | **Poster Session** — YIA G/F

■ **13:30-15:10**

Nonstationary Time Series: Past, Present and Beyond

DL12 (YIA G/F LT2)

Sponsor: Hong Kong
Chair: Wai Keung Li (The University of Hong Kong)

Nonstationary Time Series: Past, Present, and Beyond
Distinguished Lecturer: Ngai Hang Chan (The Chinese University of Hong Kong)

Invited papers:

1. **Multiple-output Quantile Regression: a Survey**
   Marc Hallin (Université Libre de Bruxelles)
2. **Bootstrap Unit Root Inference for Non-stationary Linear Processes driven by Infinite Variance Innovations**
   Giuseppe Cavaliere (University of Bologna)

Recent Advances in Large-Scale Inference

IP05 (YIA G/F LT3)

Sponsor: IMS
Organizer: Zheng Tracy Ke (The University of Chicago)
Chair: Yingying Wei (The Chinese University of Hong Kong)

Invited papers:

1. **High Dimensional Minimum Variance Portfolio Estimation**
   Yingying Li (The Hong Kong University of Science and Technology)
2. **Detecting Rare and Weak Spikes in Large Covariance Matrices**
   Zheng Tracy Ke (The University of Chicago)
3. **Neyman-Pearson Classification under High-dimensional Settings**
   Yang Feng (Columbia University)
4. **A CLT for Random Sesquilinear Forms with Applications in RMT**
   Zhonggen Su (Zhejiang University)

Recent Advances in Bayesian Functional Data Analysis

IP19 (YIA 2/F LT4)

Sponsor: Korea
Organizer: Taeryon Choi (Korea University)
Chair: Taeryon Choi (Korea University)

Invited papers:

1. **Functional Regression Approximate Bayesian Computation for Gaussian Process Density Estimation**
   David John Nott (National University of Singapore)
2. **Clustering Functional Data using Principal Curve Methods**
   Bo Wang (University of Leicester)
3. **Bayesian Spectral Analysis Models for Functional Clustering**
   Minjung Kyung (Duksung Women's University)
4. **Smoothing and Mean-covariance Estimation of Functional Data with a Bayesian Hierarchical Model**
   Dennis Cox (Rice University)

Recent Advances in Design of Experiments

IP23 (YIA 2/F LT5)

Sponsor: Chinese Society of Probability and Statistics
Organizer: Min-Qian Liu (Nankai University)
Chair: Chongqi Zhang (Guangzhou University)

Invited papers:

1. **Experimental Designs for Radar Countermeasure Reconnaissance Equipment**
   Yu Tang (Soochow University)
2. **Column-orthogonal Designs and Orthogonal Latin Hypercube Designs with Multi-dimensional Stratification**
   Jian-Feng Yang (Nankai University)
3. **Sensitivity Analysis for Computer Experiments using Permutations**
   Shifeng Xiong (Chinese Academy of Sciences)

Biostatistics and Biomedical Statistics

IP25 (YIA 2/F LT6)

Sponsor: Chinese Statistical Association (Taiwan)
Organizer: Tsung-Chi Cheng (National Chengchi University)
Chair: Jeng-Min Chiou (Academia Sinica)

Invited papers:

1. **An Additive-multiplicative Hazard Regression Model with Longitudinal Covariates**
   Yi-Kuan Tseng (National Central University)
2. **Conducting Non-experimental Comparative Research for Cancer Treatments using Large-scale Database**
   Yi-Hsin Yang (Kaohsiung Medical University)
3. **Robust Diagnostics for Multivariate Data with a Mixture of Continuous and Categorical Variables**
   Tsung-Chi Cheng (National Chengchi University)
4. **Gene Set Correlation Analysis**
   Chen-An Tsai (National Taiwan University)

Recent Advances on Algebraic Methods in Statistics

IP30 (YIA 2/F LT7)

Sponsor: Japan
Organizer: Satoshi Aoki (Kobe University)
Chair: Satoshi Kuriki (The Institute of Statistical Mathematics)

Invited papers:

1. **Sampling Methods of Fractional Factorial Designs**
   Satoshi Aoki (Kobe University)
2. **Markov Bases for Logit Models with Some Designs**
   Hisayuki Hara (Niigata University)
3. **Loading Monotonicity of Weighted Insurance Premiums, and Total Positivity Properties of Weight Functions**
   Donald Richards (The Pennsylvania State University)
4. **The Role of Algebraic Statistics in Estimation and Modeling of Random Graphs and Networks**
   Sonja Petrovic (Illinois Institute of Technology)

Modern Statistical Methods for Complex Data

IP56 (YIA 2/F LT8)

Organizer: Haipeng Shen (The University of Hong Kong)
Chair: Haipeng Shen (The University of Hong Kong)

Invited papers:

1. **Are Your Co-variates Fixed Constants or Random Variables? – It Matters**
   Lawrence David Brown (The Wharton School, University of Pennsylvania)
2. **Additive Models for Functional Data**
   Byeong Park (Seoul National University)
3. **A Simple and Practical Approach Towards Testing Global Restrictions on General Functions**
   Valentin Patilea (ENSAI)
4. **Bilinear Regression with Matrix Covariates in High Dimensions**
   Dan Yang (Rutgers University)

Confidence Interval in Some Models and High Dimensional Data Analysis

TCP21 (YIA 2/F LT9)

Organizer: Junlong Zhao (Beijing Normal University)
Chair: Chi Tim Ng (Chonnam National University)

Topic contributed papers:

1. **Confidence Distribution Inferences on the Common Value in Nonparametric Model**
   Xuhua Liu (China Agricultural University)
2. **A New Hierarchical Classification Model with Special Consideration on the Underlying Data Generating Process**
   Xiaoning Wang (Renmin University of China)
3. **A New Confidence Interval in the Errors-in-variables Model**
   Liang Yan (Beijing Institute of Technology)
4. **Conditional Expection Improved Estimation for High Dimensional SUR model**
   Li Zhao (Beijing Institute of Technology)

Numerical Computation Methods for Multivariate Distributions and Parameter Estimation

TCP22 (YIA 2/F 201)

Organizer: Masahiro Kuroda (Okayama University of Science)
Chair: Yuichi Mori (Okayama University of Science)

Topic contributed papers:

1. **Fast Estimation using the EM Algorithm for Gaussian Mixture Models**
   Masahiro Kuroda (Okayama University of Science)
2. **Approximation to the Joint Density of Eigenvalues of a Complex Wishart Matrix**
   Tatsuya Kuwabara (Tokyo University of Science)
3. **Graphical Method on an Omunibus test for Normality**
   Shigekazu Nakagawa (Kurashiki University of Science and the Arts)
4. **Bootstrapping Distributions on the Eigenvalues of Covariance Matrix**
   Hiroki Hashiguchi (Tokyo University of Science)
5. **The Application of the Mathematical Model for Fashions in Human Societies**
   Yasushi Ota (Doshisha University)

Statistical Analysis of Survival and Functional Data

TCP20 (YIA 4/F 403)

Organizer: Xin Yuan Song (The Chinese University of Hong Kong)
Chair: Jingheng Cai (Sun Yat-sen University)

Topic contributed papers:

1. **Functional Clustering of Mouse Ultrasonic Vocalization Data**
   Xiaoling Dou (Waseda University)
2. **Bayesian Joint Modeling for Survival Data with Latent Variables**
   Deng Pan (Huazhong University of Science and Technology)
3. **Bayesian Adaptive Lasso for Additive Hazard Model in Current Status Data**
   Chunjie Wang (Changchun University of Technology)
4. **Transformation Model for Sparse Functional Data**
   Guochang Wang (College of Economics)

Machine Learning Session 4

CP04 (YIA 4/F 405)

Chair: Phillip Sheung Chi Yam (The Chinese University of Hong Kong)

Contributed papers:

1. **Robust Regression under Heterogeneous Contamination**
   Hironori Fujisawa (The Institute of Statistical Mathematics)
2. **A Matrix-intensive Formulation of Factor Analysis with Specific Factors Dissociated From Errors**
   Kohei Adachi (Osaka University)
3. **Orthogonal Non-negative Matrix Tri-factorization Based on the Tweedie Family**
   Hiroyasu Abe (Doshisha University)
4. **A Factor-adjusted Multiple Testing Procedure with Application to Mutual Fund Selection**
   Lilun Du (The Hong Kong University of Science and Technology)

■ **15:10-15:30** | Coffee Break YIA G/F & 2/F

■ **15:30-17:10**

Model Selection in High-Dimensional Regression and Graphical Models and Its Applications

DL06 (YIA G/F LT2)

Sponsor: ICSA
Chair: Jin-Ting Zhang (National University of Singapore)

A New Approach to Test-based Variable Selection
Distinguished Lecturer: Tze Leung Lai (Stanford University)

Invited papers:

1. **Model Selection for High-dimensional Time Series**
   Ching-Kang Ing (Academia Sinica)
2. **Positive Definiteness of High Dimensional Regularized Covariance Matrix Estimator**
   Johan Lim (Seoul National University)

Recent Advances in Big Data Inference

IP06 (YIA G/F LT3)

Sponsor: IMS
Organizer: Yingying Fan (University of Southern California)
Chair: Daoji Li (University of Central Florida)

Invited papers:

1. **Trade-offs in Statistical Learning**
   Quentin Berthet (University of Cambridge)
2. **Optimal Correlation Detection with Application to Colocalization Analysis in Dual-channel Florescence Microscopic Imaging**
   Ming Yuan (University of Wisconsin-Madison)
3. **Robust Covariance Matrix Estimation via Matrix Depth**
   Zhao Ren (University of Pittsburgh)
4. **Neyman-Pearson (NP) Classification Algorithms and NP Receiver Operating Characteristic (NP-ROC) Curves**
   Xin Tong (University of Southern California)

Time Series and Econometrics

IP27 (YIA 2/F LT4)

Sponsor: Hong Kong
Organizer: Guodong Li and Philip L.H. Yu (The University of Hong Kong)
Chair: Guodong Li (The University of Hong Kong)

Invited papers:

1. **Testing for Stability of the Mean of Heteroskedastic Time Series**
   Liudas Giraitis (Queen Mary University of London)
2. **Linear Double Autoregressive Time Series Model and its Conditional Quantile Inference**
   Qianqian Zhu (The University of Hong Kong)
3. **COBra: Copula-based Portfolio Optimization**
   Marc Paolella (University of Zurich)
4. **Generalized Poisson Autoregressive Models for Time Series of Counts**
   Cathy W.S. Chen (Feng Chia University)

Education of Data Science and Statistics

IP29 (YIA 2/F LT5)

Sponsor: Local Host
Organizer: Helen Meng (The Chinese University of Hong Kong)
Chair: Qi-Man Shao (The Chinese University of Hong Kong)

Invited papers:

1. **The Avoidance of Data Contamination in Big Data Collection Processes**
   C.K. Wong (iASPEC Technologies Group)
2. **Educating the Next Generation of Data Scientists**
   Helen Meng (The Chinese University of Hong Kong)
3. **Recent Advances in Transfer Learning**
   Qiang Yang (The Hong Kong University of Science and Technology)
4. **From Big Data to Precision Medicine: The Role of Statisticians**
   Feifang Hu (The George Washington University)

Recent Advances in Functional Data Analysis

IP24 (YIA 2/F LT6)

Sponsor: Chinese Statistical Association (Taiwan)
Organizer: Jeng-Min Chiou (Academia Sinica)
Chair: Yi-Kuan Tseng (National Central University)

Invited papers:

1. **Modeling Functional Data Vectors**
   Jeng-Min Chiou (Academia Sinica)
2. **Sensible Functional Linear Discriminant Analysis**
   Ci-Ren Jiang (Academia Sinica)
3. **Asymptotic Perfect Discrimination of Functional Data by Penalized Discriminant Analysis**
   Lu-Hung Chen (National Chung Hsing University)
4. **Supervised Regularized Principal Component Analysis**
   Haipeng Shen (The University of Hong Kong)

Recent Advances in Non- and Semi-Parametric Inference IP31 (YIA 2/F LT7)

Sponsor: Japan
Organizer: Yoshihiko Maesono (Kyushu University)
Chair: Toshio Honda (Hitotsubashi University)

Invited papers:

1. **Statistical Estimation of Composite Risk Functionals and Risk Optimization Problems**
   Spiridon Ivanov Penev (The University of New South Wales)
2. **Some Boundary-bias-free Density Estimators**
   Yoshihide Kakizawa (Hokkaido University)
3. **Testing Symmetry of Unknown Densities via Smoothing with the Generalized Gamma Kernels**
   Masayuki Hirukawa (Setsunan University)
4. **Asymptotic Properties of Kernel Type Estimators of Ratios**
   Taku Moriyama (Kyushu University)

Recent Topics in Medical Statistics TCP24 (YIA 2/F LT8)

Organizer: Shu-Hui Chang (National Taiwan University)
Chair: Shu-Hui Chang (National Taiwan University)

Topic contributed papers:

1. **Confidence Intervals for the Difference Between Two Median Survival Times for Clustered Survival Data**
   Yu-Mei Chang (Tunghai University)
2. **Statistical Inference on Censored Data for Targeted Clinical Trials under Enrichment Design**
   Chen-Fang Chen (National Taiwan University)
3. **Dynamic Survival Prediction Using Marker Processes**
   Deng-Huang Su (Far-Eastern Polyclinics)
4. **An Adaptive Procedure to Construct Robust Genetic Association Tests using Case-parents Triad Family Data**
   Jiun-Yi Wang (Asia University)

Recent Developments in Biostatistics and Spatial TCP27 (YIA 2/F LT9)

Organizer: Yingan Liu (Nanjing Forestry University)
Chair: Junhao Pan (Sun Yat-sen University)

Topic contributed papers:

1. **The Weighting Bi-level Penalized Method for the Identification of Rare and Common Variants in Genetic Associations Studies**
   Jianwei Gou (Nanjing Forestry University)
2. **Geospatial Analysis Between Green Infrastructure Distribution and Infectious Disease Datasets**
   Jing Shen (Nanjing Forestry University)
3. **Robust Inferences for Latent Variable Model Mixed with Hidden Markov Model**
   Yemao Xia (Nanjing Forestry University)
4. **Block-based Association Tests for Rare Variants Using Kullback-Leibler Divergence**
   Degang Zhu (Nanjing Forestry University)

Recent Developments in Multiple Comparison Procedures

TCP28 (YIA 2/F 201)

Organizer: Siu Hung Cheung (The Chinese University of Hong Kong)
Chair: Mohammad Arashi (Shahrood University of Technology)

Topic contributed papers:

1. **Superiority and Non-inferiority Tests in Clinical Studies with Multiple Experimental Treatments**
   Siu Hung Cheung (The Chinese University of Hong Kong)
2. **Simultaneous Confidence Intervals for Several Quantiles of an Unknown Distribution**
   Anthony Hayter (University of Denver)
3. **Latent Ordinal Regression Models with Applications in Multiple Comparisons**
   Tong-Yu Lu (China Jiliang University)
4. **Statistical Calibration and Exact One-sided Simultaneous Tolerance Intervals for Polynomial Regression**
   Ping Yang (The Chinese University of Hong Kong)

New Statistical Methods for Graphs and Networks

TCP31 (YIA 4/F 403)

Organizer: Swati Chandna (University College London)
Chair: Yi Yu (University of Cambridge)

Topic contributed papers:

1. **Hierarchical Models for Independence Structures of Networks**
   Kayvan Sadeghi (University of Cambridge)
2. **Data Analysis using Curvature of Data Spaces and their Metric Cones**
   Kei Kobayashi (The Institute of Statistical Mathematics)
3. **Nonparametric Graphon Estimation with Covariates**
   Swati Chandna (University College London)
4. **How Many Communities Are There?**
   Diego Franco Saldana (Dow Jones)

Data Order Structure Session 2

CP15 (YIA 4/F 405)

Chair: Zhengxiao Wu (Singapore Management University)

Contributed papers:

1. **Simultaneous Confidence Bands for the Distribution Function of a Finite Population in Stratified Sampling**
   Lijie Gu (Soochow University)
2. **Blocking in Partially Replicated Two-level Factorial Designs**
   Shin-Fu Tsai (National Taiwan University)
3. **Some Results on Multi-level Factorial Designs in Complex Coding**
   Mitsunori Ogawa (Tokyo Metropolitan University)

Conference Banquet Reception & Dinner
(Science Park, ClubONE on the Park) [tickets required]

Address: Shop061-066, G/F, Building 12W, No.12 Science Park West Avenue, Hong Kong Science Park, Shatin, N.T.

- Coaches will start departing at the entrance of YIA during 17:15-18:00.

- Transportation from ClubONE on the Park to University MTR Station will be provided after the banquet.

■ **17:30-19:00** | Conference Banquet Reception (Science Park, ClubONE on the Park)

■ **19:00-21:30** | Conference Banquet Dinner (Science Park, ClubONE on the Park)

Day 3 Wed, June 29

■ 08:30-10:10

Fusion Learning - Fusing Inferences from Diverse Sources

DL03 (YIA G/F LT2)

Sponsor: IMS
Chair: Mengling Liu (NYU School of Medicine)

Invited papers:

1. **Confidence Distribution - An Effective Tool for Statistical Inference and Fusion Learning**
   Min-ge Xie (Rutgers University)
2. **Fusion Learning for Key Comparisons**
   Jan Hannig (The University of North Carolina at Chapel Hill)

Statistical Fusion Learning: Combining Inferences from Multiple Sources for More Powerful Findings
Distinguished Lecturer: Regina Y. Liu (Rutgers University)

Statistical Analysis of Dynamic Models

IP04 (YIA G/F LT3)

Sponsor: IMS
Organizer: Samuel Kou (Harvard University)
Chair: Siu Hung Cheung (The Chinese University of Hong Kong)

Invited papers:

1. **Intrinsic Noise in Nonlinear Gene Regulation Inference**
   Chao Du (University of Virginia)
2. **Efficient Parameter Inference for Dynamic Systems**
   Samuel W.K. Wong (University of Florida)
3. **Segmentation of Change-point Models**
   David Siegmund (Stanford University)
4. **Fast Analysis of Dynamic Systems via Gaussian Emulator**
   Samuel Kou (Harvard University)

Statistical Advances of Image Analysis and Spatial Statistics

IP08 (YIA 2/F LT4)

Sponsor: IMS
Organizer: Zhengjun Zhang (University of Wisconsin-Madison)
Chair: Xueqin Wang (Sun Yat-Sen University)

Invited papers:

1. **Neuron Network Detection**
   Chunming Zhang (University of Wisconsin-Madison)
2. **A Multi-resolution Scheme for Analysis of Brain Connectivity Networks**
   Vikas Singh (University of Wisconsin-Madison)
3. **Infill Asymptotics for Multivariate Spatial Processes**
   Hao Zhang (Purdue University)
4. **Study on Spatial Extreme Dependence Patterns in China's Smog Extreme Co-movements**
   Zhengjun Zhang (University of Wisconsin-Madison)

Recent Advances in Analysis of Complex Observational Data

IP42 (YIA 2/F LT5)

Organizer: Kevin He (University of Michigan)
Chair: Yi Li (University of Michigan)

Invited papers:

1. **Investigating Reliability of Diagnostic Classifications using Category-specific Measures: Lessons from Multi-center Clinical Research Networks**
   J. Richard Landis (University of Pennsylvania)
2. **Modeling Time-varying Effect of Treatment Switching with Application to Estimating the Effect of Change in Dialysis Vascular Access**
   Yuedong Wang (University of California, Santa Barbara)
3. **Big Data Regression and Prediction in Functional Genomics**
   Hongkai Ji (Johns Hopkins University)
4. **Nonparametric Estimation of Conditional Moments with Right-censored Selection Biased Data**
   Cedric Heuchenne (University of Liège)

Development in Statistical Methods for the Analysis of Events

IP10 (YIA 2/F LT6)

Sponsor: ICSA
Organizer: Zhezhen Jin (Columbia University)
Chair: Wenyang Zhang (University of York)

Invited papers:

1. **Yakovlev Promotion Time Cure Model with Local Polynomial Estimation**
   Li-Shan Huang (National Tsing Hua University)
2. **Analysis of Stratified Mark-specific Proportional Hazards Models under Two-phase Sampling with Application to HIV Vaccine Efficacy Trials**
   Yanqing Sun (The University of North Carolina at Charlotte)
3. **Nonparametric Inference for Current Status Data**
   Xingqiu Zhao (The Hong Kong Polytechnic University)
4. **Estimation of Concordance Probability with Censored Regression Models**
   Zhezhen Jin (Columbia University)

Stochastic Interacting Systems

IP13 (YIA 2/F LT7)

Sponsor: IISA
Organizer: Sunder Sethuraman (The University of Arizona)
Chair: Rabi Bhattacharya (The University of Arizona)

Invited papers:

1. **Long Range Exclusion Processes**
   Cédric Bernardin (University of Nice)
2. **Two Approximations of Coupled KPZ Equations**
   Tadahisa Funaki (The University of Tokyo)
3. **Moderate Deviation Principles for Weakly Interacting Particle Systems**
   Amarjit Budhiraja (The University of North Carolina at Chapel Hill)
4. **On KPZ-Burgers Equation and Stochastic Particle Systems**
   Sunder Sethuraman (The University of Arizona)

New Developments in Statistical Genomics

IP55 (YIA 2/F LT8)

Organizer: Yingying Wei (The Chinese University of Hong Kong)
Chair: Yingying Wei (The Chinese University of Hong Kong)

Invited papers:

1. **A Statistical Approach to Colocalizing Genetic Risk Variants in Multiple GWAS**
   Can Yang (Hong Kong Baptist University)
2. **Estimating and Accounting for Tumor Purity in Methylation Microarray Analysis**
   Hao Wu (Emory University)
3. **Spatial Temporal Modeling of Gene Expression Dynamics During Human Brain Development**
   Hongyu Zhao (Yale University)
4. **Robust Identification of Gene-environment Interactions**
   Shuangge Ma (Yale University)

Recent Advances in Bayesian Nonparametrics

TCP01 (YIA 2/F LT9)

Organizer: Subhashis Ghoshal (North Carolina State University)
Chair: Jie Hu (Xiamen University)

Topic contributed papers:

1. **Needles and Straw in a Haystack: Empirical Bayes Confidence for Possibly Sparse Sequences**
   Eduard Belitser (VU University Amsterdam)
2. **Bayesian Variable Selection for Semi-parametric Regression Models**
   Weining Shen (University of California, Irvine)
3. **An Online Gibbs Sampler Algorithm for Hierarchical Dirichlet Processes Prior**
   Yongdai Kim (Seoul National University)
4. **Bayesian Methods for Boundary Detection in Images**
   Subhashis Ghoshal (North Carolina State University)

Variable Selection and Applications of Semiparametric Models

TCP04 (YIA 2/F 201)

Organizer: Dewei Wang (University of South Carolina)
Chair: Dewei Wang (University of South Carolina)

Topic contributed papers:

1. **Embracing Blessing of Dimensionality in Factor Models**
   Quefeng Li (The University of North Carolina at Chapel Hill)
2. **Predictive Variable Selection for High-dimensional Response and Covariate**
   Yuan Wang (Washington State University)
3. **Quantile Regression in Survival Analysis with High Dimensional Sata**
   Qi Zheng (University of Louisville)
4. **Latent Variable Augmented Sparse Regression**
   Zemin Zheng (University of Science and Technology of China)

Model Selection and Hypothesis Testing

TCP32 (YIA 4/F 403)

Organizer: Jingheng Cai (Sun Yat-sen University)
Chair: Jingheng Cai (Sun Yat-sen University)

Topic contributed papers:

1. **Wild Bootstrap Tests for Serial Correlation of Time Series Objects**
   Taewook Lee (Hankuk University of Foreign Studies)
2. **Comparison of Non-nested Models Under Composite Kullback-Leibler Divergence**
   Chi Tim Ng (Chonnam National University)
3. **A Criterion-based Model Comparison Statistic for Structural Equation models with Heterogeneous Data**
   Junhao Pan (Sun Yat-sen University)
4. **Testing for the Equality of Integration Orders of Multiple Series**
   Man Wang (Donghua University)

Bayesian Session 1

CP06 (YIA 4/F 405)

Chair: Yan Liu (Ocean University of China)

Contributed papers:

1. **Bayesian Predictive Distributions in Nonparametric Function Prediction**
   Keisuke Yano (The University of Tokyo)
2. **Approximate Bayesian Inference with Pseudo-likelihood**
   Ray S W Chung (The Hong Kong University of Science and Technology)
3. **Quantile Regression-based Bayesian Nonlinear Mixed-effects Joint Models for Survival-longitudinal Data with Multiple Features**
   Yangxin Huang (University of South Florida)
4. **Bayesian Spatial Temporal Model for Air Pollutant Data**
   Wai Ming Li (The Hong Kong University of Science and Technology)
5. **A Copula Based Multivariate Hierarchical Spatial Model with Applications to Daily Air Pollutant Extremes in Pearl River Delta**
   Ka Shing Chan (The Hong Kong University of Science and Technology)

New Developments in High-Dimensional Spatial and Spatio-Temporal Modeling

IP41 (YIA 5/F 505)

Organizer: Chae Young Lim (Seoul National University)
Chair: Chae Young Lim (Seoul National University)

Invited papers:

1. **High Dimensional Variable Selection for Spatial Regression**
   Taps Maiti (Michigan State University)
2. **Computational Challenges with Big Environmental Data**
   Marc G. Genton (King Abdullah University of Science and Technology)
3. **Hierarchical Low Rank Approximation for Large Spatial Datasets**
   Ying Sun (King Abdullah University of Science and Technology)
4. **Spatial Methods for Nonstationary Fields Using Compact Basis Functions**
   Soutir Bandyopadhyay (Lehigh University)

■ 10:10-10:30 | Coffee Break

YIA G/F & 2/F

■ 10:30-12:10

Heterogeneity in Large-Scale Data, with Connections to Causal Inference

DL08 (YIA G/F LT2)

Sponsor: Bernoulli Society
Chair: Regina Y. Liu (Rutgers University)

The Power of Heterogeneous Large-scale Data for High-dimensional Causal Inference
Distinguished Lecturer: Peter Lukas Buhlmann (ETH, Zurich, Seminar for Statistics)

Invited papers:

1. **Penalized Estimation Methods for High-dimensional Causal Discovery**
   Ali Shojaie (University of Washington)
2. **Goodness of Fit Tests for High-dimensional Linear Models**
   Rajen Dinesh Shah (University of Cambridge)

Variational Inference

DL16 (YIA G/F LT3)

Sponsor: Australia and New Zealand
Chair: Alan Welsh (Australian National University)

Fast Approximate Inference for Arbitrarily Large Statistical Models via Message Passing
Distinguished Lecturer: Matt Wand (University of Technology Sydney)

Invited papers:

1. **Variational Approximations for Directional Data of Arbitrary Dimension**
   Jay Breidt (Colorado State University)
2. **Flexible Online Multivariate Regression with Variational Bayes and the Matrix-variate Dirichlet Process**
   Meng Hwee Victor Ong (National University of Singapore)

Nonparametric Modelling and High Dimensional Data Analysis

IP11 (YIA 2/F LT4)

Sponsor: ICSA
Organizer: Wenyang Zhang (University of York)
Chair: Zhezhen Jin (Columbia University)

Invited papers:

1. **Empirical Bayes Prediction for the Multivariate Newsvendor Loss Function**
   Gourab Mukherjee (University of Southern California)
2. **Low-dimensional Confounder Adjustment and High-dimensional Penalized Estimation for Survival Analysis**
   Jialiang Li (National University of Singapore)
3. **Innovated Interaction Screening for High-dimensional Nonlinear Classification**
   Daoji Li (University of Central Florida)
4. **High-dimensional A-learning for Optimal Dynamic Treatment Regimes**
   Rui Song (North Carolina State University)

Recent Advances in Complex Data Analysis

IP22 (YIA 2/F LT5)

Sponsor: Chinese Society of Probability and Statistics
Organizer: Liping Zhu (Renmin University of China)
Chair: Zhongyi Zhu (Fudan University)

Invited papers:

1. **Empirical Likelihood Inference in Linear Regression with Nonignorable Missing Response**
   Wangli Xu (Renmin University of China)
2. **On Marginal Sliced Inverse Regression for Ultrahigh Dimensional Model-free Feature Selection**
   Zhou Yu (East China Normal University)
3. **Ensemble Sufficient Dimension Folding Methods on Analyzing Matrix-valued Data**
   Yuan Xue (University of International Business and Economics)
4. **Efficient and Nonparametric Causal Inference**
   Zheng Zhang (Renmin University of China)

Statistical Methodology for Biomedical Sciences

IP43 (YIA 2/F LT6)

Organizer: Xuming He (University of Michigan)
Chair: Mi-Ok Kim (Cincinnati Children's Hospital Medical Center)

Invited papers:

1. **Approximate Median Regression for Complex Survey Data with Skewed Response**
   Stuart Lipsitz (Harvard University)
2. **Identification of Homogeneous and Heterogeneous Variables in Pooled Cohort Studies**
   Mengling Liu (NYU School of Medicine)
3. **Conditional Graphical Models with Applications in Integrative Genomics**
   Jie Peng (University of California, Davis)
4. **Test for Genomic Imprinting Effects on the X Chromosome**
   Wing Kam Fung (The University of Hong Kong)

Challenges and Recent Advances in Methods for Missing Data Problems

IP52 (YIA 2/F LT7)

Organizer: Zonghui Hu (Biostatistics Research Branch, DCR)
Chair: Chiung-Yu Huang (Johns Hopkins University)

Invited papers:

1. **Improved Estimation of Average Treatment Effects on the Treated: Local Efficiency, Double Robustness, and Beyond**
   Zhiqiang Tan (Rutgers University)
2. **Bayesian Pattern Mixture Models for the Analysis of Repeated Attempt Designs**
   Michael Daniels (The University of Texas at Austin)
3. **Semiparametric Pseudoscore Estimation for Regressions with Potentially High-dimensional but Incompletely Observed Covariatesa**
   Zonghui Hu (National Institutes of Health)
4. **Nonparametric Sufficient Dimension Reduction with Missing Predictors at Random**
   Qihua Wang (Chinese Academy of Sciences)

New Statistical Methods for Genetic Data Analysis

TCP11 (YIA 2/F LT8)

Organizer: Anne Buu (University of Michigan)
Chair: Runze Li (The Pennsylvania State University)

Topic contributed papers:

1. **Robust Modeling of RNA-Seq Data**
   Hui Jiang (University of Michigan)
2. **Random Field Modelling of Genetic Association for Sequencing Data in Family-based Studies**
   Ming Li (Indiana University)
3. **The Genetic Architecture of Complex Phenotypes: New Insight from Game Theory**
   Rongling Wu (The Pennsylvania State University)
4. **Estimation of Stratified Mark-specific Proportional Hazards Models under Two-phase Sampling with Application to HIV Vaccine Efficacy Trials**
   Guangren Yang (Jinan University)
5. **An Efficient Genome-wide Association Test for Multivariate Phenotypes Based on the Fisher Combination Function**
   James Yang (University of Michigan)

Statistical Issues in Analyzing Bioinformatics Data

TCP26 (YIA 2/F LT9)

Organizer: Xiaodan Fan (The Chinese University of Hong Kong)
Chair: Han Li (Shenzhen University)

Topic contributed papers:

1. **Lengthening and Shortening of Tumour-derived Plasma DNA: Reconciling a Long-standing Controversy**
   Peiyong Jiang (The Chinese University of Hong Kong)
2. **Estimating Reproducibility in Genome-wide Association Studies**
   Weichuan Yu (The Hong Kong University of Science and Technology)
3. **DNA Methylation in Enhancers**
   Jiangwen Zhang (The University of Hong Kong)
4. **Tumor Purity and DMR Estimation from DNA Methylation Data**
   Xiaoqi Zheng (Shanghai Normal University)

Recent Advances on Random Processes and Related Problems

TCP29 (YIA 2/F 201)

Organizer: Zhonggen Su (Zhejiang University)
Chair: Zhonggen Su (Zhejiang University)

Topic contributed papers:

1. **The Fourth Moment Theorem for the Complex Multiple Wiener-Itô Integrals**
   Yong Liu (Peking University)
2. **Volume Growth and Escape Rate of Symmetric Diffusion Processes**
   Shun-Xiang Ouyang (South University of Science and Technology of China)
3. **Double Contour Integral Formulas in Two Matrix Model and Related Non-intersecting Brownian Motions**
   Dong Wang (National University of Singapore)
4. **The Stochastic Logarithmic Schrödinger Equation**
   Deng Zhang (Shanghai Jiao Tong University)
5. **Metric Entropy of High Dimensional Convex Functions**
   Fuchang (Frank) Gao (University of Idaho)

Statistical Inference Under Model Uncertainties

TCP30 (YIA 4/F 403)

Organizer: Stephen Lee (The University of Hong Kong)
Chair: Stephen Lee (The University of Hong Kong)

Topic contributed papers:

1. **On a General Procedure for Constructing Confidence Sets under Partially Identified Models**
   Han Jiang (The University of Hong Kong)
2. **Variable Selection in Single-index Varying Coefficient Models**
   Anna Liu (University of Massachusetts Amherst)
3. **Quadratic Discriminant Analysis for High-dimensional Data**
   Yilei Wu (University of Waterloo)
4. **Spherical Cap Packing Asymptotics and Rank-extreme Detection**
   Kai Zhang (The University of North Carolina at Chapel Hill)

Probability Session 1

CP08 (YIA 4/F 405)

Chair: Kouji Tahata (Tokyo University of Science)

Contributed papers:

1. **A New Two-sample Test for High-dimension, Low-sample-size Data**
   Aki Ishii (University of Tsukuba)
2. **Modified Variational Mode Decomposition using Ebayesthresh**
   Guebin Choi (Seoul National University)
3. **The Kumaraswamy Skew G Distributions**
   Rui Li (The University of Manchester)
4. **On Moment Based Density Approximations for Aggregate Losses**
   Jeffrey Chu (The University of Manchester)

■ **12:10-13:30** | Lunch — Chung Chi Tang Student Canteen

■ **13:45-17:00** | **Excursion [tickets required]**

Route 1: Hong Kong City Tour (The Peak, Repulse Bay, Stanley and Murray House)
Route 2: Hong Kong Cultural Tour (Ngong Ping Piazza, Tian Tan Buddha, Po Lin Monastery and Wisdom Path)
- Coaches will start departing at the entrance of YIA during 13:45-14:00

Day 4 Thu, June 30

■ 08:30-10:10

Recent Advances in Machine Learning for Personalized Medicine DL01 (YIA G/F LT2)

Sponsor: IMS
Chair: Wenbin Lu (North Carolina State University)

Recent Advances in Machine Learning for Personalized Medicine
Distinguished Lecturer: Michael Kosorok (The University of North Carolina at Chapel Hill)

Invited papers:

1. **Tree-based Method for High-dimensional Survival Data**
   Ruoqing Zhu (University of Illinois at Urbana-Champaign)
2. **How Many Processors Do We Really Need in Parallel Computing?**
   Guang Cheng (Purdue University)

Random Matrices and High-Dimensional Statistics IP26 (YIA G/F LT3)

Sponsor: Hong Kong
Organizer: Jianfeng Yao (The University of Hong Kong)
Chair: Debashis Paul (University of California, Davis)

Invited papers:

1. **Free Probability and High Dimensional Time Series**
   Arup Bose (Indian Statistical Institute)
2. **A Universal High-dimensional Data Structural Detection Approach via Random Matrix Theory**
   Guangming Pan (Nanyang Technological University)
3. **Extreme Eigenvalues of Large-dimensional Spiked Fisher Matrices with Application**
   Qinwen Wang (University of Pennsylvania)
4. **Homoscedasticity Tests Valid in Both Low and High-dimensional Regressions**
   Jianfeng Yao (The University of Hong Kong)

Adaptive Randomization in Clinical Trials IP09 (YIA 2/F LT4)

Sponsor: IMS
Organizer: Feifang Hu (The George Washington University)
Chair: Feifang Hu (The George Washington University)

Invited papers:

1. **Adaptive Multi-arm Platform Designs for Screening Effective Treatments via Predictive Probability**
   J. Jack Lee (The University of Texas MD Anderson Cancer Center)
2. **Randomization in Small Population Clinical Trials**
   William Rosenberger (George Mason University)
3. **Central Limit Theorems of a Recursive Stochastic Algorithm with Applications to Adaptive Designs**
   Li-Xin Zhang (Zhejiang University)
4. **Nonparametric Response Adaptive Randomization Procedures Based on p-values**
   Zhongqiang Liu (Henan Polytechnic University)

Recent Developments in the Analysis of High-Dimensional Time Series with Nonstationarities

IP40 (YIA 2/F LT5)

Organizer: Haeran Cho (University of Bristol)
Chair: Young K. Lee (Kangwon National University)

Invited papers:

1. **Shrinkage Estimation for Multivariate Hidden Markov Models**
   Mark Fiecas (The University of Warwick)
2. **Sparse High-dimensional Varying Coefficient Models**
   Eun Ryung Lee (Sungkyunkwan University)
3. **Non-stationary Dynamic Factor Models for Large Datasets**
   Matteo Barigozzi (London School of Economics and Political Science)
4. **Change-point Detection in High-dimensional Panel Data**
   Haeran Cho (University of Bristol)

Network Models: Theory and Methods

IP14 (YIA 2/F LT6)

Sponsor: IISA
Organizer: Shankar Bhamidi (The University of North Carolina at Chapel Hill)
Chair: Shankar Bhamidi (The University of North Carolina at Chapel Hill)

Invited papers:

1. **Geometry of Random Graphs: Scaling Limits and Universality**
   Sanchayan Sen (TU Eindhoven)
2. **Dynamic Causal Networks with Multi-scale Temporal Structure**
   Eric Kolaczyk (Boston University)
3. **Dense and Sparse Graph Limits Arising from Respondent Driven Sampling**
   Adrian Roellin (National University of Singapore)

Statistical Inferences for Complex Data

IP36 (YIA 2/F LT7)

Sponsor: Singapore
Organizer: Jin-Ting Zhang (National University of Singapore)
Chair: Gourab Mukherjee (University of Southern California)

Invited papers:

1. **High-dimensional Linear Hypothesis Testing Under Heteroscedasticity**
   Jin-Ting Zhang (National University of Singapore)
2. **Modelling Liquidity Supply in Limit Order Book with a Vector Functional Autoregressive (VFAR) model**
   Ying Chen (National University of Singapore)
3. **Principal Flows and Sub-manifolds**
   Zhigang Yao (National University of Singapore)
4. **Density Estimation in the Two-sample Problem with Likelihood Ratio Ordering**
   Tao Yu (National University of Singapore)

Advanced Modeling of Large-Scale Dependent Data

TCP02 (YIA 2/F LT8)

Organizer: Hiroki Masuda (Kyushu University)
Chair: Hiroki Masuda (Kyushu University)

Topic contributed papers:

1. **Statistical Inferences for Ergodic Point Processes and Application to Limit Order Books**
   Simon Clinet (The University of Tokyo)
2. **Statistical Inference for Price Discovery: a Stochastic Process Approach**
   Yuta Koike (Tokyo Metropolitan University)
3. **Parametric Inference for Diffusion Processes with High-frequency Financial Data**
   Teppei Ogihara (The Institute of Statistical Mathematics)
4. **Mighty Convergence in Mixed-rates Asymptotics**
   Yusuke Shimizu (Kyushu University)

Recent Developments in Large and Complex Data Analysis

TCP25 (YIA 2/F LT9)

Organizer: Xinghao Qiao (London School of Economics and Political Science)
Chair: Shaojun Guo (Renmin University of China)

Topic contributed papers:

1. **Nonlinear Shrinkage Estimation of Large Integrated Covariance Matrix**
   Qilin Hu (London School of Economics and Political Science)
2. **Generalised Additive and Index Models with Shape Constraints**
   Yining Chen (London School of Economics and Political Science)
3. **Spatial Weight Matrix Estimation in a Dynamic Spatial Autoregression Model**
   Cheng Qian (London School of Economics and Political Science)
4. **Multi-zoom Autoregressive Models**
   Rafal Baranowski (London School of Economics and Political Science)

Statistical Modeling and Its Applications

TCP34 (YIA 2/F 201)

Organizer: Xuejun Jiang (South University of Science and Technology of China)
Chair: Xuejun Jiang (South University of Science and Technology of China)

Topic contributed papers:

1. **Detection of Gene Regulatory Relationships by ODE Model with Time Lags**
   Jie Hu (Xiamen University)
2. **An Extended Mallows' Model for Rank Data Aggregation with Covariates**
   Han Li (Shenzhen University)
3. **Spatial Analysis of Water Quality in Fujian Bay, China**
   Yan Liu (Ocean University of China)
4. **The Latent Low Rank Model to Colocalize Genetic Risk Variants in Multiple GWAS**
   Jin Liu (Duke-NUS Medical School)

Statistical Modeling in Economics and Finance

TCP35 (YIA 4/F 403)

Organizer: Yan Liu (Ocean University of China)
Chair: Xin Zhao (Ocean University of China)

Topic contributed papers:

1. **Energy Consumption and Economic Growth: Evidence from Coastal areas in China**
   Jing Guo (Ocean University of China)
2. **Empirical Research on the Market Volatility of Copper Future of SHFE**
   Ruifen Huang (Ocean University of China)
3. **Research on Marine Economy Efficiency and its influencing Factors in the Blue Economic Region of China**
   Yong Peng (Ocean University of China)
4. **What Matters More? Developing an Integrated Weighting Technique for Coastal Vulnerability to Storm Surge**
   Shun Yuan (Ocean University of China)
5. **Does the Relationship Between Insurance Development and Economic Growth Maintain Stability? An Empirical Analysis of Coastal Area in China Based on Non-parametric Local Polynomial Regression**
   Hui Zheng (Ocean University of China)

Probability Session 2

CP09 (YIA 4/F 405)

Chair: Phillip Sheung Chi Yam (The Chinese University of Hong Kong)

Contributed papers:

1. **Favourable Non-extreme Region: Outlier**
   Tudzla Hernita (STIS 53 Computer Division)
2. **Randomly Weighted Sums of Conditionally Dependent Random Variables with Applications to Risk Theory**
   Dongya Cheng (Soochow University)
3. **Precise Local Large Deviations for Random Sums with Applications to Risk Models**
   Fengyang Cheng (Soochow University)
4. **Kolmogorov-type Inequality and Some Limit Theorems for Extended Negatively Dependent Random Variables**
   Jigao Yan (Soochow University)

■ **10:10-10:30** | Coffee Break — YIA G/F & 2/F

■ 10:30-12:10

Particle Representations for Measure-Valued Processes and Stochastic Partial Differential Equations

DL05 (YIA G/F LT2)

Sponsor: IMS
Chair: Sunder Sethuraman (University of Arizona)

Particle Representations for Stochastic Partial Differential Equations
Distinguished Lecturer: Thomas G. Kurtz (University of Wisconsin-Madison)

Invited papers:

1. **SPDE with Poisson Representation**
   Jie Xiong (University of Macau)
2. **Uniform in Time Particle System Approximations for Nonlinear Equations of Keller-Segel Type**
   Wai Tong Fan (University of Wisconsin-Madison)

Random Fields: Theory and Applications

IP45 (YIA 2/F LT4)

Organizer: Yimin Xiao (Michigan State University)
Chair: Zhengjun Zhang (University of Wisconsin-Madison)

Invited papers:

1. **Local Times of Gaussian Random Fields on the Sphere**
   Xiaohong Lan (University of Science and Technology of China)
2. **Gaussian Random Fields with Stationary Increments and their Asymptotic Properties**
   Wensheng Wang (Hangzhou Normal University)
3. **Wavelet Estimators of Multivariable Nonparametric Regression Functions with Long Memory Data**
   Dongsheng Wu (The University of Alabama in Huntsville)
4. **Estimation of Fractal Indices for Bivariate Gaussian Random Fields**
   Yimin Xiao (Michigan State University)

Emerging Statistical Methods in Big Data Analytics

IP49 (YIA 2/F LT5)

Organizer: Ping Ma (University of Georgia)
Chair: Hongkai Ji (Johns Hopkins University)

Invited papers:

1. **Pooling Partial Observations for Efficient Estimation of the Joint Distribution**
   Xiaodan Fan (The Chinese University of Hong Kong)
2. **Automated Feature Identification using Online Knowledge and EMR Data**
   Sheng Yu (Tsinghua University)
3. **Discovering Association Patterns via Theme Dictionary Models**
   Ke Deng (Tsinghua University)
4. **Impact of Genotyping Errors on Statistical Power of Association Tests in Genomic Analyses**
   Lin Hou (Tsinghua University)

Emerging Nonparametric Methods for Financial Data

IP47 (YIA 2/F LT6)

Organizer: Lan Xue (Oregon State University)
Chair: Fang Li (Indiana University-Purdue University Indianapolis)

Invited papers:

1. **High-dimensional Inference with an Application to Financial Network**
   Yongli Zhang (University of Oregon)
2. **Spline Confidence Bands for Generalized Regression Models**
   Jing Wang (University of Illinois at Chicago)
3. **Monotone Additive Models in Productivity Analysis**
   Lan Xue (Oregon State University)
4. **Oracally Efficient Estimation and Consistent Model Selection for ARMA Time Series with Trend**
   Lijian Yang (Soochow University)

Advances in Analysis of Complex High-Dimensional Data

IP59 (YIA 2/F LT7)

Organizer: Shaojun Guo (Renmin University of China)
Chair: Xin Tong (University of Southern California)

Invited papers:

1. **Estimation of Nonsmooth Functionals**
   Mark Low (University of Pennsylvania)
2. **Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity**
   Tony Cai (University of Pennsylvania)
3. **Valid Inference on Semiparametric Estimators with Regressors Generated by High Dimensional Regularization**
   Shaojun Guo (Renmin University of China)
4. **High Dimensional and Functional Autoregressions with an Application to Functional Volatility Processes**
   Xinghao Qiao (London School of Economics and Political Science)

Recent Advances in Biomarker Evaluation and Risk Prediction

TCP09 (YIA 2/F LT8)

Organizer: Shanshan Li (Indiana University)
Chair: Ruoqing Zhu (University of Illinois Urbana-Champaign)

Topic contributed papers:

1. **Variable Selection and Predictive Performance for Correlated Biomarkers using Regularized Regression Approaches**
   Wei-Ting Hwang (University of Pennsylvania)
2. **Estimation of Covariate-specific Time-dependent ROC Curves in the Presence of Missing Biomarkers**
   Shanshan Li (Indiana University)
3. **Personalized Differential Gene Expression Detection**
   Yingying Wei (The Chinese University of Hong Kong)
4. **Combine Longitudinal Biomarkers in Predicting Time-dependent Risks**
   Hong Zhu (University of Texas Southwestern Medical Center)

Advanced Bayesian Modeling

TCP33 (YIA 2/F LT9)

Organizer: Junhao Pan (Sun Yat-sen University)
Chair: Junhao Pan (Sun Yat-sen University)

Topic contributed papers:

1. **Bayesian Analysis of the Functional-coefficient Autoregressive Heteroscedastic Model**
   Jingheng Cai (Sun Yat-sen University)
2. **Bayesian Adaptive Lasso for Multivariate Generalized Linear Model with Latent Variables**
   Xiang-Nan Feng (The Chinese University of Hong Kong)
3. **Bayesian Approaches in Analyzing Earthquake Catastrophic**
   Xuejun Jiang (South University of Science and Technology of China)
4. **Bayesian Transformation Quantile Regression**
   Pengfei Liu (Jiangsu Normal University)
5. **Efficient Transformations for Exploring MCMC Sampler on a Family of Banana-shaped Distributions**
   Maolin Pan (Nanjing University)

Recent Advances in High Dimensional Estimation Theory

TCP07 (YIA 2/F 201)

Organizer: Gourab Mukherjee (University of Southern California)
Chair: Gourab Mukherjee (University of Southern California)

Topic contributed papers:

1. **Bayes Projection and its Applications to High-dimensional Problems**
   Fumiyasu Komaki (The University of Tokyo)
2. **A Nonparametric Bayesian Approach for Sparse Sequence Estimation**
   Yunbo Ouyang (University of Illinois at Urbana-Champaign)
3. **Optimal Community Detection in Degree-corrected Stochastic Block Models**
   Anderson Ye Zhang (Yale University)
4. **Group-linear Empirical Bayes Estimation of a Heteroscedastic Normal Mean**
   Asaf Weinstein (Stanford University)

High-Dimensional Data Analyses with Application in Biomedical Studies

TCP05 (YIA 4/F 403)

Organizer: Yi Li (University of Michigan)
Chair: Ming Gao Gu (The Chinese University of Hong Kong)

Topic contributed papers:

1. **Covariance-insured Screening Methods for Ultrahigh Dimensional Variable Selection**
   Kevin (Zhi) He (University of Michigan)
2. **Detecting Association to Precision Networks via Conditional Multi-type Graphical Models**
   Yanming Li (University of Michigan)
3. **Detecting Rare and Faint Signals via Thresholding Maximum Likelihood Estimators**
   Yumou Qiu (University of Nebraska–Lincoln)
4. **Automatic Detection of Significant Areas for Functional Data with Directional Error Control**
   Peirong Xu (Southeast University)
5. **Set-based Test for Gene-environment Interaction**
   Baqun Zhang (Renmin University of China)

Big Data/Network Session 1

CP11 (YIA 4/F 405)

Chair: Xuexia Wang (University of Wisconsin-Milwaukee)

Contributed papers:

1. **Monitoring the Results of Cardiac Surgery Based on 3 or More Outcomes by Variable Life-adjusted Display**
   Fah Fatt Gan (National University of Singapore)
2. **An Adaptive 3-D Image Denoising Framework**
   Partha Sarathi Mukherjee (Boise State University)
3. **New Types of Shrinkage Estimators of Poisson Means**
   Genso(Yuan-Tsung) Watanabe(Chang) (Mejiro University)
4. **Spatial Autoregressive Model Estimation for Large-scale Social Networks**
   Yingying Ma (Beihang University)
5. **Finite Sample Covariance Estimation Based on Graphical Model**
   Chaojie Wang (The Chinese University of Hong Kong)

■ **12:10-13:30** | Lunch — Chung Chi Tang Student Canteen

■ **13:00-13:45** | **Poster Session** — YIA G/F

■ **13:30-15:10**

From Cells to Populations: Modeling and Inference for Genomic Data

DL04 (YIA G/F LT2)

Sponsor: IMS

Chair: Susan Wilson (The University of New South Wales and The Australian National University)

Modelling and Inference of Co-ancestry in Populations

Distinguished Lecturer: Elizabeth Thompson (University of Washington)

Invited papers:

1. **Three Dimensional Chromatin Structure and Spatial Gene Regulation**
   Shili Lin (The Ohio State University)
2. **Advances of Bayesian Nonparametrics in Population Genetics of Infectious Diseases**
   Vladimir Minin (University of Washington)

Random Networks

DL15 (YIA G/F LT3)

Sponsor: India
Chair: Arup Bose (Indian Statistical Institute)

Drainage Networks and the Brownian Web

Distinguished Lecturer: Rahul Roy (Indian Statistical Institute)

Invited papers:

1. **Rumor Spreading in Dynamic Networks**
   Marco Isopi (Sapienza University of Rome)
2. **Application of Random Graphs in Epidemiology and Economics**
   Farkhondeh Alsadat Sajadi (University of Isfahan)

Recent Advances in Lifetime Data Analysis

IP48 (YIA 2/F LT4)

Organizer: Mei-Ling Ting Lee (University of Maryland)
Chair: Xingqiu Zhao (The Hong Kong Polytechnic University)

Invited papers:

1. **Causal Inference and Time**
   Odd O. Aalen (University of Oslo)
2. **Survival and Quality of Life**
   Catherine Louise Huber-Carol (Université Paris Descartes)
3. **Nonparametric Analysis of the Dependence Structure for Recurrent Gap Time Data**
   Shu-Hui Chang (National Taiwan University)
4. **Estimating Model-based Attributable Risk Functions in the Asian Cohort Consortium**
   Ying Qing Chen (Fred Hutchinson Cancer Research Center)

Recent Developments in Survival Analysis and Personalized Medicine

IP46 (YIA 2/F LT5)

Organizer: Wenbin Lu (North Carolina State University)
Chair: Rui Song (North Carolina State University)

Invited papers:

1. **Concordance-assisted Learning for Estimating Optimal Individualized Treatment Regimes**
   Wenbin Lu (North Carolina State University)
2. **Semiparametric Structural Equation Models with Latent Variables for Right-censored Data**
   Kin Yau Wong (The University of North Carolina at Chapel Hill)
3. **Flexible Modeling of Bivariate Recurrent Events Data**
   Limin Peng (Emory University)
4. **A Semiparametrically Efficient Estimator of the Time-varying Effects for Survival Data with Time-dependent Treatment***
   Huazhen Lin (Southwestern University of Finance and Economics)

Some Recent Developments in High-Frequency Financial Econometrics

IP53 (YIA 2/F LT6)

Organizer: Bingyi Jing (The Hong Kong University of Science and Technology)
Chair: Bingyi Jing (The Hong Kong University of Science and Technology)

Invited papers:

1. **Testing the Equality of Large U-statistic Based Correlation Matrices**
   Xinsheng Zhang (Fudan University)
2. **On the Integrated Systematic and Idiosyncratic Volatility with Large Panel High-frequency Data**
   Xinbing Kong (Soochow University)
3. **Testing for Presence of Leverage Effect Under High Frequency**
   Zhi Liu (University of Macau)
4. **Adaptive Thresholding for Large Volatility Matrix Estimation Based on High-frequency Financial Data**
   Cuixia Li (Lanzhou University)

Analysis of Spatial and Spatio-Temporal Data

IP33 (YIA 2/F LT7)

Sponsor: India
Organizer: Debasis Sengupta (Indian Statistical Institute)
Chair: Marc G. Genton (King Abdullah University of Science and Technology)

Invited papers:

1. **Modeling Tangential Vector Fields on the Sphere**
   Debashis Paul (University of California, Davis)
2. **A Novel Nonparametric Threshold-free Method to Produce Functional MRI Activation Maps**
   Rajesh Ranjan Nandy (UNT Health Science Center)
3. **A Statistical Description of the Spatial Extent of a Spell of Rainfall**
   Subrata Kundu (The George Washington University)
4. **Zero Expectile Processes and Bayesian Spatial Regression**
   Anandamayee Majumdar (Soochow University)

Big Data/Network Session 2

CP12 (YIA 2/F LT9)

Chair: Fah Fatt Gan (National University of Singapore)

Contributed papers:

1. **Enhanced Construction of Gene Regulatory Networks using Hub Gene Information**
   Donghyeon Yu (Keimyung University)
2. **A Popularity Scaled Latent Space Model for Network Structure Formulation**
   Xiangyu Chang (Xi'an Jiaotong University)
3. **Network Dynamics Detection using Liquid Association**
   Tianwei Yu (Emory University)
4. **Novel Nonparametric Methods to Test Rare Variants for Multiple Traits**
   Xuexia Wang (University of Wisconsin-Milwaukee)

Probability Session 3

CP10 (YIA 2/F 201)

Chair: Zheng Zhang (Renmin University of China)

Contributed papers:

1. **On Measure of Second-order Marginal Symmetry for Multi-way Contingency Tables**
   Yusuke Saigusa (Tokyo University of Science)
2. **Asymptotic Normality of Naive Canonical Correlation Coefficient in High Dimension Low Sample Size**
   Mitsuru Tamatani (Doshisha University)
3. **Generalized Asymmetry Models and Separations of Symmetry for Square Tables**
   Kouji Tahata (Tokyo University of Science)
4. **Log-normal Distribution Type Symmetry Model for Ordinal Square Contingency Tables**
   Kiyotaka Iki (Tokyo University of Science)

Data Order Structure Session 3

CP16 (YIA 4/F 403)

Chair: Chun Yip Yau (The Chinese University of Hong Kong)

Contributed papers:

1. **GARCH Modeling of Five Popular Commodities**
   Stephen Chan (The University of Manchester)
2. **Decomposing Time Series into Oscillation Components with Random Frequency Modulation**
   Takeru Matsuda (The University of Tokyo)
3. **A New Approach for Analyzing Panel AR(1) Series with Application to the Unit Root Test**
   Yu-Pin Hu (National Chi Nan University)
4. **Detecting Differentially Methylated Regions via Non-homogeneous Hidden Markov Model**
   Linghao Shen (The Chinese University of Hong Kong)
5. **Bootstrap Method for Autoregressive Model**
   Bambang Suprihatin (University of Sriwijaya)

Data Order Structure Session 4

CP17 (YIA 4/F 405)

Chair: Leming Qu (Boise State University)

Contributed papers:

1. **Modified Kaplan-Meier Estimator and Nelson-Aalen Estimator with Geographical Weighting for Survival Data**
   Guanyu Hu (Florida State University)
2. **Model Selection of Switching Mechanism for Financial Time Series**
   Chau Buu Truong (Feng Chia University)
3. **Adjusted Kaplan-Meier Survival Curves for Marginal Treatment Effect in Observational Studies**
   Xiaofei Wang (Duke University School of Medicine)
4. **Quantile Regression Based on A Weighted Approach under Semi-competing Risks Data**
   Jin-Jian Hsieh (National Chung Cheng University)
5. **Regression Modeling of Interval Censored Cometing Risk Data with Missing Cause**
   YangJin Kim (Sookmyung Women's University)

■ **15:10-15:30** | Coffee Break

YIA G/F & 2/F

■ 15:30-17:10

Statistical Inference for Stochastic Processes: Asymptotic Theory and Implementation

DL14 (YIA G/F LT2)

Sponsor: Japan
Chair: Hiroki Masuda (Kyushu University)

Statistics for Stochastic Processes: Inferential and Probabilistic Aspects
Distinguished Lecturer: Nakahiro Yoshida (The University of Tokyo)

Invited papers:

1. **Hybrid Type Estimation for Diffusion Type Processes Based on High Frequency Data**
   Masayuki Uchida (Osaka University)
2. **Computational Aspects of Simulation and Inference for CARMA and COGARCH Models**
   Stefano Iacus (The University of Milan)

Recent Advances and Trends in Time Series Analysis

IP17 (YIA G/F LT3)

Sponsor: Bernoulli Society
Organizer: Liudas Giraitis (Queen Mary University of London)
Chair: Liudas Giraitis (Queen Mary University of London)

Invited papers:

1. **On Consistency/Inconsistency of MDL Model Selection for Piecewise Autoregressions**
   Richard A. Davis (Columbia University)
2. **Structure Identification in Panel Data Analysis**
   Wenyang Zhang (University of York)
3. **Root-n Consistent Estimation of the Marginal Density of Some Stationary Time Series**
   Lionel Truquet (ENSAI)
4. **Inference for Conditionally Heteroscedastic Location-scale Time Series Models**
   Sangyeol Lee (Seoul National University)

Advances in Statistical Inference for Multivariate Response Data

IP39 (YIA 2/F LT4)

Sponsor: Australia and New Zealand
Organizer: Samuel Mueller (University of Sydney)
Chair: Samuel Mueller (University of Sydney)

Invited papers:

1. **Vector Regression Without Marginal Distributions or Association Structures**
   Alan Huang (The University of Queensland)
2. **Estimation and Inference in Directional Mixed Models for Compositional Datas**
   Janice Lea Scealy (Australian National University)
3. **Bayesian Gegenbauer Long Memory Financial Time Series Models**
   Jennifer Chan (The University of Sydney)
4. **Stability Selection with Information Criteria**
   Zhen Pang (The Hong Kong Polytechnic University)

Semiparametric Statistical Methods for Complex Problems

IP50 (YIA 2/F LT6)

Organizer: Lan Wang (University of Minnesota)
Chair: Lan Wang (University of Minnesota)

Invited papers:

1. **Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection**
   Runze Li (The Pennsylvania State University)
2. **Modeling Complex and Big Survival Data: Computation and More**
   Yi Li (University of Michigan)
3. **Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference**
   Ingrid van Keilegom (Université catholique de Louvain)
4. **Semi-nonparametric Inference in Possibly Misspecified Regression Models with Missing Data**
   Phillip Sheung Chi Yam (The Chinese University of Hong Kong)

Stochastic Partial Differential Equations

IP58 (YIA 2/F LT7)

Organizer: Jian Song (The University of Hong Kong)
Chair: Jian Song (The University of Hong Kong)

Invited papers:

1. **Density of Parabolic Anderson Field**
   Yaozhong Hu (The University of Kansas)
2. **Poincare Inequality for Dirichlet Distributions and Infinite-dimensional Generalizations**
   Feng-Yu Wang (Beijing Normal University)
3. **Large Deviation Principle for the Occupation Measures of Stochastic PDEs**
   Lihu Xu (University of Macau)

Bayesian Session 2

CP07 (YIA 2/F LT8)

Chair: Tianwei Yu (Emory University)

Contributed papers:

1. **New Method for Revealing Free Energy Landscape of Proteins**
   Hangjin Jiang (The Chinese University of Hong Kong)
2. **A Fast and Powerful W-test for Pairwise Gene-gene Interaction Testing in GWAS Data**
   Maggie Wang (The Chinese University of Hong Kong)
3. **Closed Form Bayesian Inferences for Binary Logistic Regression**
   Kevin Dayaratna (The Heritage Foundation)

Machine Learning Session 5

CP05 (YIA 2/F LT9)

Chair: Hironori Fujisawa (The Institute of Statistical Mathematics)

Contributed papers:

1. **A One-stage Approach for Principal Component Regression via L1-type Regularization**
   Shuichi Kawano (The University of Electro-Communications)
2. **Generalized Principal Component Analysis: Dimensionality Reduction through the Projection of Natural Parameters**
   Yoonkyung Lee (The Ohio State University)
3. **A High Dimensional Two-sample Test using Nearest Neighbors Based on a New Dissimilarity Measure**
   Rahul Biswas (Indian Statistical Institute)

Big Data/Network Session 3

CP13 (YIA 2/F 201)

Chair: Zheng Zhang (Renmin University of China)

Contributed papers:

1. **A Dynamic Logistic Regression for Network Link Prediction**
   Jing Zhou (Peking University)
2. **Promote Gross Capital Formation with Internet Access**
   Aulia Dini (STIS 53 Economics Division)
3. **A Cognitive Model of Data Visualization in Big Data Analytics**
   Ken W Li (Hong Kong Institute of Vocational Education)
4. **The Estimation of Claims Reserve with Reserving by Detailed Conditioning Method (RDC)**
   Adhitya Ronnie Effendie (Gadjah Mada University)

Data Order Structure Session 5

CP18 (YIA 4/F 403)

Chair: Yu-Pin Hu (National Chi Nan University)

Contributed papers:

1. **Comparing Predictive Values of Two Diagnostic Tests in Small Clinical Trials**
   Kouji Yamamoto (Osaka University)
2. **On Coding and Centering in the Autologistic Regression Model**
   Mark Wolters (Fudan University)
3. **Automatic Optimal Batch Size Selection for Recursive Estimators of Time-average Covariance Matrix**
   Kin Wai Chan (Harvard University)
4. **Covariate-adjusted Response-adaptive Designs for Weibull Survival Responses**
   Ayon Mukherjee (Queen Mary University of London)

Data Order Structure Session 6

CP19 (YIA 4/F 405)

Chair: YangJin Kim (Sookmyung Women's University)

Contributed papers:

1. **The Empirical Beta Copula**
   Hideatsu Tsukahara (Seijo University)
2. **Multivariate Copula Density Estimation by Mixture of Parametric Copula Densities**
   Leming Qu (Boise State University)
3. **Nonparametric Wind Power Forecasting under Fixed and Random Censoring**
   Georgios Effraimidis (University of Southern California)
4. **Dynamic Prediction of Alzheimer's Disease Risk Based on Longitudinal Biomarkers and Functional Data**
   Sheng Luo (University of Texas at Houston)

Poster Session

Venue: G/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong (CUHK)

Birth Cohort Effect in Japan - Automatic Detection and Statistical Evaluation Tetsuji Tonda (Prefectural University of Hiroshima)	PS02
High Dimensional LASSO Variable Selection Under Strongly-correlated Covariates Kaimeng Zhang (Chonnam National University)	PS03
Growth Curve Model with Nonparametric Baselines and Its Statistical Inference Kenichi Satoh (Hiroshima University)	PS04
Evaluating Standard Errors of Total Heritability Estimate in Genome-wide Association Studies Based on Summary Statistics Alone Hon-Cheong So (The Chinese University of Hong Kong)	PS05
Estimating Regression Coefficients Including Nuisance Baseline and Its Applications Kenichi Kamo (Sapporo Medical University)	PS06
Joint Inference for a GLMM Model with a NLME Covariate Model Subject to Left Censoring and Measurement Error, with Application to AIDS Studies Hongbin Zhang (City University of New York)	PS07
A New Distribution To Describe Big Data Yuanyuan Zhang (The University Of Manchester)	PS09
Nonlinear Operator Estimation with Bayes Sieve Estimator Masaaki Imaizumi (The University of Tokyo)	PS10
On the Spectral Distribution of Hayashi's Estimator for High Dimensional Stock Price Process Arnab Chakrabarti (Indian Statistical Institute)	PS11
Recent Advances in Approximate Solution for Stochastic Differential Delay Equation Young-Ho Kim (Changwon National University)	PS12
Unified Tree-structured Non-crossing Quantile Regression Model Jaeoh Kim (Korea University)	PS13

Day 1
Mon, June 27

Abstracts

Mon, June 27 (10:00-11:00) l PL01

Understanding Importance Sampling

Plenary Speaker: Persi Diaconis (Stanford University)

Chair: Qi-Man Shao (The Chinese University of Hong Kong)

Understanding Importance Sampling

Persi Diaconis

Stanford University, United States. *diaconis@math.stanford.edu*

Abstract: Importance sampling is a mainstay of scientific computing giving us the ability to change one (easy to generate) probability distribution into another (hard to generate). In practice, this often results in very long tailed samples where one observation can dominate a million others. The usual way of evaluating the accuracy of importance sampling uses the variance. This is a poor idea for long tailed data. In joint work with Sourav Chatterjee we introduce a different criteria which is (a) necessary and sufficient for accuracy (b) 'easy' to estimate in natural problems.

In the talk, I will review the many varied uses of importance sampling (in things like rare event simulation and particle filters), introduce our new criteria, show that monitoring the empirical variance 'on the fly' can be really deceptive and give some practical examples from sequential importance sampling to show the new criteria in action.

Mon, June 27 (11:10-12:10) l PL02

Low Rank Structure in Highly Multivariate Models

Plenary Speaker: Iain Johnstone (Stanford University)

Chair: Ming-Yen Cheng (National Taiwan University)

Low Rank Structure in Highly Multivariate Models

Iain Johnstone

Stanford University, United States. *imj@stanford.edu*

Abstract: In 1964 Alan James gave a remarkable classification of many of the eigenvalue distribution problems of multivariate statistics. We show how the classification readily adapts to contemporary 'spiked models' -- high dimensional data with low rank structure. In particular we approximate likelihood ratios when the number of variables grows proportionately with sample size or degrees of freedom. High dimensions bring phase transition phenomena, with quite different likelihood ratio behavior for small and large spike strengths. James' framework allows a unified approach to problems such as signal detection, matrix denoising, regression and canonical correlations.

Mon, June 27 (13:30-15:10) l DL07 l Sponsor: IISA

Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics

Chair: Amarjit Budhiraja (The University of North Carolina at Chapel Hill)

Distinguished Lecturer: Rabi Bhattacharya (The University of Arizona)

Analysis of Non-Euclidean Data: Use of Differential Geometry in Statistics

Rabi Bhattacharya

The University of Arizona, United States. *rabi@math.arizona.edu*

Abstract: This talk focuses on examples of (1) differential geometric depictions of certain classes of digital images arising in biology, medicine, machine vision and other fields of science and engineering and (2) their model-independent statistical analysis for purposes of identification, discrimination and diagnostics. In addition to commonly known Euclidean submanifolds such as spheres with important implications for tectonics, as examples we mention D.G. Kendall's landmarks based shape spaces, certain graphical models for evolutionary biology, and the space of 3x3 diffusion matrices arising in diffusion tensor imaging. Nonparametric statistical inference based on Fre'chet means as minimizers of expected squared distances have been recently used effectively for such non-Euclidean spaces M, such as manifolds. We discuss general issues of uniqueness of the Fre'chet minimizer and the consistency and asymptotic distribution of its empirical estimate under these distances. Applications in medicine and neuroscience include discrimination between a normal organ and a diseased one in the human body for the diagnosis of glaucoma and certain types of schizophrenia, and changes in the geometric structure of the white matter in the brain's cortex brought about by Parkinson's disease, Alzheimer's, schizophrenia, autism, etc.

Research supported by the NSF Grant DMS 1406872

Dimension Reduction on Tori and Polyspheres

Stephan Huckemann

University Göttingen, Germany. *huckeman@math.uni-goettingen.de*

Abstract: Although tori as being compact, zero curvature manifolds seem to be comparatively simple and well behaved, finding a rich sequence of lower dimensional subspaces for the purpose of dimension reduction is challenging.

Tangent space approximations cannot take into account periodicity and even worse, intrinsic approaches are completely inapplicable because, for instance, almost all geodesics are dense.

In the approach proposed in this talk, after adaptively changing the geometry into a stratified spherical one (with singularities and identifications), a variant of principal nested spheres analysis by Jung et al. (2012) becomes available. It turns out that this torus-PCA extends from tori (products of circles) to polysphere-PCA (on products of spheres of arbitrary dimension).

Applying torus-PCA to RNA structure data as well as polysphere-PCA to medial representations of body organs shows that this new approach goes well beyond previous analyses.

Nonparametric Regression on Manifolds

Lizhen Lin

The University of Texas at Austin, United States. *lizhen.lin@austin.utexas.edu*

Abstract: Over the last few decades data represented in various non-conventional forms have become increasingly more prevalent. Typical examples include diffusion matrices in diffusion tensor imaging (DTI) of neuroimaging, and various digital imaging data. One may also encounter data that are stored in the forms of subspaces, orthonormal frames, surfaces, and networks. Statistical analysis of such data requires rigorous formulation and characterization of the underlying space, and inference is heavily dependent on the geometry of the space. For a majority of the cases considered, the underlying spaces where these general data objects lie on, fall into the general category of manifolds. This talk focuses on nonparametric regression on manifolds where either responses or predictors are on the manifolds. In particular, we present extrinsic local regression models for regressions with manifold valued responses, and construct Gaussian process models for regression and classification with manifold valued predictors.

Mon, June 27 (13:30-15:10) l DL09 l Sponsor: Korea

Recent Advances in Covariance Estimation

Chair: Woncheol Jang (Seoul National University)

Distinguished Lecturer: Ja-Yong Koo (Korea University)

Covariance Estimation with the Positive Definite Constraint

Ja-Yong Koo

Korea University, South Korea. *jykoo@korea.ac.kr*

Abstract: We develop a nonparametric estimation of a symmetric positive-definite matrix response given covariates. We first derive a lower bound on the rate of convergence under the assumption that the Kullback-Leibler divergence between the conditional distributions for the response given covariates has a quadratic bound. This lower bound technique based on the Cholesky decomposition is of independent interest and can be used for other conditional matrix estimation problems. We then establish that this lower bound is actually a minimax optimal rate of convergence.

Local Distance Regression Model for Manifold-valued Data

Hongtu Zhu

The University of North Carolina at Chapel Hill, United States. *htzhu@email.unc.edu*

Abstract: The talk is to introduce a local distance regression model for the analysis of manifold-valued response and its association with multiple covariates of interest, such as age or gender, in Euclidean space. Such manifold-valued data arises frequently in medical imaging, surface modeling, and computer vision, among many others. We develop an intrinsic distance regression model solely based on an moment assumption, avoiding specifying any parametric distribution and projecting data to local tangent planes. We characterize nonlinear association between the Euclidean space of multiple covariates and manifold-valued responses. We develop an estimation procedure to calculate the parameter estimates and determine their asymptotic distributions. We construct the local and global test statistics to test hypotheses of unknown parameters. Simulation studies and a real data analysis are used to evaluate the finite sample properties of our methods.

Solution Path of Condition-number Regularization

Joong-Ho Won

Seoul National University, South Korea. *wonj@stats.snu.ac.kr*

Abstract: The recently introduced condition-number-regularized covariance estimation method (CondReg) has been demonstrated to be highly useful for estimating high-dimensional covariance matrices. Unlike L1-regularized estimators, this approach has the added advantage that no sparsity assumptions are made. The regularization path of the lasso solution has received much attention in the literature. Despite their importance, the solution paths of covariance estimators however have not been considered in much detail. In this paper, we provide a complete characterization of the entire solution path of the CondReg estimator. Our characterization of the solution path has important applications as it yields fast algorithms that compute the CondReg estimates for all possible values of the regularization parameter at the same cost as that for a single fixed parameter. We present two instances of fast algorithms: the forward and the backward algorithms. These algorithms greatly speed up the cross-validation procedure that selects the optimal regularization parameter. Our new method is efficiently implemented with the R package CondReg.

Mon, June 27 (13:30-15:10) | IP01 | Sponsor: IMS

Recent Advances on Analysis of High-Dimensional Data

Organizer: Runze Li (The Pennsylvania State University)

Chair: Runze Li (The Pennsylvania State University)

De-biasing Regularized Estimators With High-dimensional Data

Cun-Hui Zhang

Rutgers University, United States. *czhang@stat.rutgers.edu*

Abstract: We consider statistical inference of smooth functions of the unknown in a semi-low-dimensional approach to the analysis of high-dimensional data. In a general setting and a number of specific examples, we discuss regular and super-efficient sample size requirements for de-biasing regularized estimators. We also discuss the benefit of unlabeled data for the estimation of linear functionals in semi supervised linear regression.

CoCoLasso for High-dimensional Error-in-variables Regression

Hui Zou

University of Minnesota, United States. *zouxx019@umn.edu*

Abstract: Much theoretical and applied work has been devoted to high-dimensional regression with clean data. However, we often face corrupted data in many applications where missing data and measurement errors cannot be ignored. Loh and Wainwright (2012, AoS) proposed an interesting non-convex modification of the Lasso for doing high-dimensional regression with noisy and missing data. The non-convexity formulation brings up the issue of multiple local minimizers. Through some careful analysis, they showed that a projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of all global minimizers. In this article, we argue that the virtues of convexity contribute fundamentally the success and popularity of the Lasso. In light of this, we propose a new method named CoCoLasso that is convex and can handle a general class of corrupted datasets including the cases of additive measurement error and random missing data. CoCoLasso automatically enjoys the benefits of convexity for high-dimensional regression. We derive the statistical error bounds of CoCoLasso as well as its sign-consistent selection property. We demonstrate the superior performance of our method over the non-convex approach in Loh and Wainwright (2012) by simulation studies.

Slow Kill for Big Data Screening

Yiyuan She

Florida State University, United States. *yshe@stat.fsu.edu*

Abstract: Modern large-scale statistical datasets may involve millions of observations and features. An efficient learning scheme is proposed by gradually removing variables based on a criterion and a schedule. The resultant algorithms build variable screening into estimation and the fact that the problem size keeps dropping throughout the iterations makes the scheme particularly suitable for big data learning. Theoretical guarantees of low statistical error are provided in the presence of design coherence. Experiments on real and synthetic data show that the proposed method compares very well with boosting and other state of the art methods in regression and classification while being computationally scalable.

This is joint work with Adrian Barbu.

Nonparametric Two-sample Test in Ultra-high Dimension

Lan Wang

University of Minnesota, United States. *wangx346@umn.edu*

Abstract: In genomics and quantitative finance, multivariate data are often heavy tailed. We propose a new nonparametric two-sample test for comparing the mean vectors from two populations in the ultra-high dimensional setting. The critical value of the test is approximated using high-dimensional multiplier bootstrap method. Our method naturally incorporates the dependence structure of the data and allows the covariance structures of the two populations to be different. The improved power performance comparing with normality-based tests is demonstrated via Monte Carlo simulations.

Mon, June 27 (13:30-15:10) | IP34 | Sponsor: India

Robust Statistical Methods

Organizer: Subhra Sankar Dhar (Indian Institute of Technology, Kanpur)

Chair: Subhra Sankar Dhar (Indian Institute of Technology, Kanpur)

Big Data World: Wide Consensus in Estimation using Parallelized Inference

Juan Antonio Cuesta-Albertos

Universidad de Cantabria, Spain. *juan.cuesta@unican.es*

Abstract: Nowadays it is not strange to see people involved with problems with such a huge amount of data, that the analysis has to be carried out in several separated units (centers, countries,...), each one providing its own estimation of the object or parameter of interest.

In this talk we will assume that the goal is to estimate the full distribution of a multidimensional random variable. Therefore, the analyst needs to find a new distribution summarizing all the estimations (distributions) received from the units. Somehow, we can say that he needs to find a consensus between the proposals he has received; however, it may be the case that some estimations are too different from the rest as to allow a full consensus to be reached. To fix this conflict, we propose a procedure to reach a not so restrictive "wide consensus" between the received estimations which tries to summarize the estimations of most of units; but not necessarily all of them.

Our development requires to handle a metric between distributions. Our selection is the Wasserstein distance. This distance enjoys very nice mathematical properties, the problem being the difficulty of its computation in dimensions higher than 1. Thus, we also propose an effective algorithm which allows to compute a consensus distribution summarizing a finite number of proposals if we restrict ourselves to fixed location-scatter families, not necessarily elliptical.

On Bootstrap and Robustness of Regularized Kernel Based Methods

Andreas Christmann

University of Bayreuth, Germany. *andreas.christmann@uni-bayreuth.de*

Abstract: Regularized empirical risk minimization plays an important role in machine learning theory. Special cases are pattern recognition, regression, quantile regression, and pairwise learning such as ranking. In this talk a broad class of regularized learning methods based on kernels and their empirical bootstrap approximations will be investigated from the viewpoints of consistency and statistical robustness. If time allows, some results on localized learning for the big data situation will be given, too.

Keywords: machine learning, kernel methods, regularized risk minimization, robustness, bootstrap.

References:

(1) Christmann, A. and Zhou, D.-X. (2015). On the Robustness of Regularized Pairwise Learning Methods Based on Kernels. Submitted. Preprint on arXiv.
(2) Steinwart, I. and Christmann, A. (2008). Support Vector Machines. Springer, New York.
(3) Christmann, A. and Hable, R. (2013). On the Consistency of the Bootstrap Approach for Support Vector Machines and Related Kernel Based Methods. Chapter 20 in: Empirical Inference. Festschrift in Honor of Vladimir N. Vapnik. Eds. B. Schoelkopf, Z. Luo, V. Vovk. Springer, New York.. pp. 231-244.
(4) Christmann, A., Salibian-Barrera, M., Van Aelst, S. (2013). Qualitative Robustness of Bootstrap Approximations for Kernel Based Methods. Chapter 16 in Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather, Eds. C. Becker, S. Kuhnt, R. Fried (2013), pp. 263-278.Springer, Heidelberg, New York.

Robust Estimation of Precision Matrices under Cellwise Contamination

Garth Tarr

The University of Newcastle, Australia. *garth.tarr@sydney.edu.au*

Abstract: There is a great need for robust techniques in data mining and machine learning contexts where many standard techniques such as principal component analysis and linear discriminant analysis are inherently susceptible to outliers. Furthermore, standard robust procedures assume that less than half the observation rows of a data matrix are contaminated, which may not be a realistic assumption when the number of observed features is large. Tarr, Müller and Weber (2016) consider the problem of estimating covariance and precision matrices under cellwise contamination. Specifically, the use of a robust pairwise covariance matrix as an input to various regularisation routines, such as the graphical lasso, QUIC and CLIME. We review a number of approaches that can be used to ensure the input covariance matrix is positive semidefinite. The result is a potentially sparse precision matrix that is resilient to moderate levels of cellwise contamination and scales well to higher dimensions. A comparison between the pairwise approach and other standard robust techniques, such as the MCD, is made in terms of entropy loss, various matrix norms and Gaussian graphical discovery rates. We briefly discuss the selection of an appropriate value for the tuning parameter that controls the level of sparsity and potential applications involving financial data and bioinformatics.

Tarr G, Müller S and Weber NC (2016). Robust estimation of precision matrices under cellwise contamination, Computational Statistics and Data Analysis, 93, 404-420. DOI: 10.1016/j.csda.2015.02.005.

Penalized Weighted Least Squares for Outlier Detection and Robust Regression

Xiaoli Gao

The University of North Carolina at Greensboro, United States. *x_gao2@uncg.edu*

Abstract: In this talk, we propose a novel approach, called "penalized weighted least squares" (PWLS), for simultaneous outlier detection and robust regression. Under some conditions, if the lasso penalty is used, the PWLS estimator of the regression coefficients is consistent and equivalent to a redescending M-estimator; if the adaptive Lasso penalty is used, the PWLS has strong robustness and is able to detect the outlier set consistently. Under sufficient conditions, the PWLS have the same asymptotic normality as the weight parameter is given in advance. The small-sample performance of the new approach is demonstrated via simulations and real applications. This is a joint work with Yixin Fang.

Mon, June 27 (13:30-15:10) l IP37 l Sponsor: Australia and New Zealand

Multivariate Smoothing

Organizer: Martin Hazelton (Massey University)

Chair: Spiridon Ivanov Penev (The University of New South Wales)

Unconstrained Multivariate Bandwidth Selection for Density and Density Derivative Estimation

Jose E. Chacon

University of Extremadura, Spain. *jechacon@unex.es*

Abstract: In this talk we review some results on recent significant progress in the field of multivariate bandwidth selection. Duong and Hazelton (2003) developed plug-in methodology to choose an unconstrained bandwidth matrix from the data. As a result, they obtained a data-driven multivariate kernel density estimator with a smoothing parameter in its most general form. However, their proposal still relied on the use of constrained pilot bandwidths, having the form of a positive multiple of the identity matrix. The first plug-in bandwidth selector using unconstrained bandwidths at all levels of smoothing was proposed in Chacón and Duong (2010), and was shown to consistently outperform all the previous approaches in practice. Chacón, Duong and Wand (2011) extended these results and studied the asymptotics of general multivariate kernel density derivative estimators, and Chacón and Duong (2013) explicitly provided data-based algorithms to choose an unconstrained bandwidth matrix for multivariate density derivative estimation.

Asymptotics of Variable-bandwidth Kerne Density and Density-ratio Estimation for Planar Point Patterns in Epidemiology

Tilman M. Davies

University of Otago, New Zealand. *tdavies@maths.otago.ac.nz*

Abstract: The two-dimensional kernel density estimator has long proved an invaluable tool for estimation of the density functions of spatially continuous point patterns. It has been shown that highly heterogeneous patterns, such as those that frequently occur in applications in geographical epidemiology, benefit a great deal from adaptive or variable-bandwidth smoothing. Such an approach aims to impart greater detail in areas with a relative abundance of data by a reduction in smoothing, while simultaneously smoothing over sparsely populated regions. The theoretical and applied benefits extend to estimation of the density-ratio function, typically recognised as the popular kernel relative risk surface. However, the problem of bandwidth selection under the adaptive framework is a difficult one, and, until now, has been complicated further by the fact a closed-form expression for the asymptotic mean integrated squared error (AMISE) for this estimator was unavailable. We derive the full expression for the asymptotic bias and hence, by building on previous work where asymptotic variance was addressed, we are finally able to write down the AMISE for both the adaptive density and density-ratio estimator. This paves the way for deeper insight into the performance of bivariate adaptive kernel smoothing and the future development of bandwidth selection guidelines for this flexible technique.

Probit Transformation for Nonparametric Kernel Estimation of the Copula Density

Gery Geenens

The University of New South Wales, Australia. *ggeenens@unsw.edu.au*

Abstract: Copula modeling has become ubiquitous in modern statistics. Here, the problem of nonparametrically estimating a copula density is addressed. Arguably the most popular nonparametric density estimator, the kernel estimator is not suitable for the unit-square-supported copula densities, mainly because it is heavily affected by boundary bias issues. In addition, most common copulas admit unbounded densities, and kernel methods are not consistent in that case. In this work, a kernel-type copula density estimator is proposed. It is based on the idea of transforming the uniform marginals of the copula density into normal distributions via the probit function, estimating the density in the transformed domain, which can be accomplished without boundary problems, and obtaining an estimate of the copula density through back-transformation. Although natural, a raw application of this procedure was, however, seen not to perform very well in the earlier literature. Here, it is shown that, if combined with local likelihood density estimation methods, the idea yields very good and easy to implement estimators, fixing boundary issues in a natural way and able to cope with unbounded copula densities. The asymptotic properties of the suggested estimators are derived, and a practical way of selecting the crucially important smoothing parameters is devised. Finally, extensive simulation studies and a real data analysis evidence their excellent performance compared to their main competitors.

Multivariate Log-density Estimation with Applications to Approximate Likelihood Inference

Martin Hazelton

Massey University, New Zealand. *M.Hazelton@massey.ac.nz*

Abstract: This talk concerns kernel estimation of the logarithm of a multivariate probability density function. The work is motivated primarily by approximate likelihood inference, in which we seek to construct an estimate of the log-likelihood function for a complex model by estimating the constituent log-densities from simulated realizations of the model. The properties of the kernel log-density estimator are heavily influenced by the unboundedness of the log function at zero. In particular, standard asymptotic expansions can provide a poor guide to finite sample behaviour for this estimator, with consequences for the choice of methodology for bandwidth selection. In response we develop a new approximate cross-validation bandwidth selector. We explore its theoretical properties and describe its finite sample behaviour through numerical experiments. This bandwidth selection methodology is then applied to estimation of log-likelihoods from model simulations. Our implementation of approximate likelihood inference is illustrated through an example in which a complex genetic model is used in estimating migration rates between village communities on the Indonesian island of Sumba.

Mon, June 27 (13:30-15:10) | IP21 | Sponsor: Chinese Society of Probability and Statistics

Complex Data Analysis

Organizer: Niansheng Tang (Yunnan University)

Chair: Liping Zhu (Renmin University of China)

Functional Single-index Model for Functional Data

Zhongyi Zhu

Fudan University, China. *zhuzy@fudan.edu.cn*

Abstract: In this paper, the functional single-index model is proposed for functional response data, possibly measured with error, for predictor and response. Using the two-step algorithm, the coefficient functions and the unknown smooth link function are estimated, and their asymptotic properties are studied under mild conditions. Simulation studies and an application of intra-day volatility patterns of the S&P 500 index are conducted to illustrate our method and theory.

On Improving Efficiency by Borrowing Information across Quantiles

Yuanyuan Lin

The Chinese University of Hong Kong, Hong Kong. *ylin@sta.cuhk.edu.hk*

Abstract: As a competitive alternative to the least squares regression, the quantile regression is a popular statistical tool for the modeling and inference of conditional quantile function. In conventional quantile regression models, major complications involve in the semiparametric efficient estimation arise from curve estimation and computational difficulty. In this talk, I will discuss a nearly efficient estimation and inference procedure for quantile regression models, as well as a new quantile regression framework, that is possibly superior to the existing methods in terms of estimation accuracy. Numerical studies with supportive evidence are presented.

Nonparametric Model for Panel Data with Fixed Effects and Locally Stationary Regressors

Tao Huang

Shanghai University of Finance and Economics, China. *huang.tao@mail.shufe.edu.cn*

Abstract: In this talk, we study a nonparametric model for panel data with fixed effects and locally stationary regressors in a setting where both the time series and the cross section are large and the nonparametric regression function changes smoothly over time. A nonparametric pooled profile likelihood method is proposed and the asymptotic results are established. Numerical studies and real data analyses are presented to illustrate the performance of the proposed method.

Mon, June 27 (13:30-15:10) I IP51

Complex Data Analysis: Theory and Methods

Organizer: Yang Feng (Columbia University)

Chair: Yang Feng (Columbia University)

Screening and Feature Selection in High-dimensional Fisher Discriminant Analysis

Ming-Yen Cheng

National Taiwan University, Taiwan. *cheng@math.ntu.edu.tw*

Abstract: Fisher discriminant analysis is a powerful tool for classification problems in many different subject areas. In addition to statistical data analysis, it is arguably one of the key elements in data mining, machine learning, and other contexts in data science. In the new age of big data, it happens often that both the dimensionality p and the sample size n are large. When the dimensionality is large, consistency, computational efficiency, and numerical stability can be very serious issues. In this case, it is often desirable to reduce the dimensionality by employing screening and variable selection methods, assuming some sparsity condition. Existing screening approaches use marginal criterion that ignores dependence between features, which is hard to check in practice. We propose a greedy screening method and establish its screening consistency property in the p>>n regime. Simulation study and a real data example are presented to justify the efficacy of the proposed approach.

Adaptive Sparse Non-linear Metric Learning via Boosting

Tian Zheng

Columbia University, United States. *tz33@columbia.edu*

Abstract: Distance measures have been foundation of various learning methods, whose efficiency heavily depends on the chosen measure. Much effort had been contributed to learning an appropriate distance metric. In a high dimensional setting, however, traditional metric learning methods face the challenge that many input variables bring in noises that mask the true signal hidden in a low-dimensional subspace, as well as resulting in a formidable computational cost. In this talk, we address these issues by adaptively learning a sparse distance metric in a high-dimensional space using a novel boosting based algorithm. Simulations and experiments with several real datasets show that our approach compares favorably with the state-of-the-art methods in the current metric learning literature.

QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization

Lucy Xia

Stanford University, United States. *lucyxia@stanford.edu*

Abstract: We propose a novel Rayleigh quotient based sparse quadratic dimension reduction method – named QUADRO (Quadratic Dimension Reduction via Rayleigh Optimization) – for analyzing high dimensional data. Unlike in the linear setting where Rayleigh quotient optimization coincides with classification, these two problems are very different under nonlinear settings. In this paper, we clarify this difference and show that Rayleigh quotient optimization may be of independent scientific interests. One major challenge of Rayleigh quotient optimization is that the variance of quadratic statistics involves all fourth cross-moments of predictors, which are infeasible to compute for high-dimensional applications and may accumulate too many stochastic errors. This issue is resolved by considering a family of elliptical models. Moreover, for heavy-tail distributions, robust estimates of mean vectors and covariance matrices are employed to guarantee uniform convergence in estimating nonpolynomially many parameters, even though only the fourth moments are assumed. Methodologically, QUADRO is based on elliptical models which allow us to formulate the Rayleigh quotient maximization as a convex optimization problem. Computationally, we propose an efficient linearized augmented Lagrangian method to solve the constrained optimization problem. Theoretically, we provide explicit rates of convergence in terms of Rayleigh quotient under both Gaussian and general elliptical models. Thorough numerical results on both synthetic and real datasets are also provided to back up our theoretical results.

Estimating Network Edge Probabilities by Neighborhood Smoothing

Ji Zhu

University of Michigan, United States. *jizhu@umich.edu*

Abstract: The problem of estimating probabilities of network edges from the observed adjacency matrix has important applications to predicting missing links and network denoising. It has usually been addressed by estimating the graphon, a function that determines the matrix of edge probabilities, but is ill-defined without strong assumptions on the network structure. Here we propose a novel computationally efficient method based on neighborhood smoothing to estimate the expectation of the adjacency matrix directly, without making the strong structural assumptions graphon estimation requires. The neighborhood smoothing method requires little tuning, has a competitive mean-squared error rate, and outperforms many benchmark methods on the task of link prediction in both simulated and real networks. This is joint work with Yuan Zhang and Elizaveta Levina.

Mon, June 27 (13:30-15:10) | TCP15

Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques II: Robust SEM, Multilevel SEM, Meta-Analytic SEM, Latent Profile Analysis, and Cross-Classified SEM

Organizer: Oi-Man Kwok (Texas A&M University)

Chair: Oi-Man Kwok (Texas A&M University)

Meta-analysis: A Structural Equation Modeling Approach

Mike W.-L. Cheung

National University of Singapore, Singapore. *mikewlcheung@nus.edu.sg*

Abstract: Meta-analysis is widely used as a research tool to synthesize research findings in many disciplines including, psychology, management, education, and medical research. Structural equation modeling (SEM) is another popular technique to test hypothesized models in the behavioral and social sciences. This talk gives a brief overview on how meta-analytic models can be formulated as structural equation models to conduct (1) fixed- and random-effects meta-analyses; (2) multivariate meta-analysis; and (3) three-level meta-analysis. Examples will be used to illustrate these models. Advantages and limitations of this approach will be discussed.

The Impact of ICC on the Effectiveness of Level-specific Fit Indices in Multilevel Structural Equation Modeling: A Monte Carlo Study

Hsien-Yuan Hsu

University of Mississippi, United States. *hsuhy0914@gmail.com*

Abstract: Several researchers have recommended level-specific fit indices should be applied to detect the lack of model fit at any level in multilevel structural equation models. While we concur with their view, we note that these studies did not sufficiently consider the impact of intraclass correlation (ICC) on the performance of level-specific fit indices and our study proposed to fill this gap in the methodological literatures. A Monte Carlo study was conducted to investigate the performance of (a) level-specific fit indices (⟦CFI⟧_(⟦PS⟧_(W/B)), ⟦TLI⟧_(⟦PS⟧_(W/B)), and ⟦RMSEA⟧_(⟦PS⟧_(W/B))) derived by partially saturated model method and (b) ⟦SRMR ⟧_W and ⟦SRMR⟧_B in terms of their performance in MSEM across varying ICCs. The design factors included intra-class correlation (ICC: ICC1=.091 to ICC6=.500), numbers of groups in between-level models (NG: 50, 100, 200, and 1000), group size (GS: 30, 50, and 100), and type of misspecification (no misspecification, between-level misspecification, and within-level misspecification). Our study raise a concern regarding the performance of between-level-specific fit indices in low ICC conditions. Simulation results suggested the performance of ⟦TLI⟧_(PS_B) and ⟦RMSEA⟧_(PS_B) were more influenced by ICC compared to ⟦CFI⟧_(PS_B) and ⟦SRMR⟧_B. However, when traditional cutoff values (RMSEA ≤ 0.06; CFI, TLI ≥ 0.95; SRMR ≤ 0.08; Hu & Bentler, 1999) were applied, ⟦CFI⟧_(PS_B) and ⟦TLI⟧_(PS_B) were still able to detect misspecified between-level models even when ICC was as low as .091 (ICC1). ⟦RMSEA⟧_(PS_B) and ⟦SRMR⟧_B are not recommended if ICC is as low as .091 (ICC1) and .231 (ICC3), respectively.

Using Multivariate-t-based Maximum Likelihood for Robust Structural Equation Modeling

Hok Chio Lai

University of Cincinnati, United States. *mark.lai@uc.edu*

Abstract: Although normal-theory maximum likelihood is commonly used in structural equation modeling (SEM) and is asymptotically efficient and consistent, it is not robust to outliers and bad leverage points. As shown in previous literature, a small amount of such data contamination can lead to erroneous results in SEM. However, as SEM models usually handle a large number of variables, it is more tedious to identify both univariate and multivariate outliers and leverage points, and more difficult to spot their presence. In conventional regression analyses, the Student's t distribution has been used to obtain robust inference even with the presence of outliers, as the t distribution with a small degree of freedom has a bigger tail than the normal distribution to allow for outliers. Recently, maximum likelihood based on the multivariate t distribution has been implemented in Mplus, and our preliminary results showed that it can produce more efficient and less biased SEM parameter estimates with the presence of outliers and leverage points. In this presentation we aim to systematically examine the robustness of the multivariate-t-distribution-based maximum likelihood to obtain consistent parameter estimates under different sample sizes and proportions and mechanisms of outliers and leverage points, using Monte Carlo simulation. We also compare the performance of this method with the alternative method of using Huber-type weight functions to downweigh the influence of outliers, and note the pros and cons of each approach. Due to its ease of use, we recommend substantive researchers to routinely use the multivariate-t-maximum likelihood to guard against the problem of data contamination.

Applying Latent Profile Analysis on the Development of Self-control and Self-esteem Configuration Among Adolescents

Yuan-Hsuan Lee

National Taichung University of Education, Taiwan. *jasvi.rms@gmail.com*

Abstract: The Latent Profile Analysis (LPA) is a probabilistic model for classifying individuals into homogeneous groups in multivariate interval data within the Structural Equation Modeling framework, such that individuals in one group are similar to each other but are different from individuals in other groups. This study applied LPA to examine the self-control and self-esteem configuration among Taiwanese adolescents. Self-control and self-esteem are two highly distinguishable traits about the "self." Self-esteem, defined as a person’s overall evaluation toward oneself, has been a widely researched topic in adolescent quality of life, academic adjustment, and interpersonal relationship. Likewise, self-control, which focuses on the regulation of the self, has greatly researched due to its correlates with student adjustment. Nevertheless, there is a lack of empirical evidence regarding how the two constructs, when put together, relate to an adolescent’s life. Using latent profile analysis (LPA), this study investigated the configuration of self-control and self-esteem among junior high school students across two consecutive years. Also examined were the effect of the self profiles on indicators of quality of life. Results of the study revealed four distinct profiles among adolescents based on their self-control and self-esteem. The profile groups included the "Quality Selves (high SC-SE)," "Disadvantageous Selves (low SC-SE)," "Baseline," and "Self-Esteem." The grouping effect on students’ deviant behavior, life satisfaction, friendship, time management, and academic achievement shed light on the intervention and prevention of student adjustment problems. Educational programs solely aim at cherishing self could move beyond for a double-core direction that also enhances adolescent social adaption with self-discipline training.

Testing Mediation Effects in Cross-classified Multilevel Data

Wen Luo

Texas A&M University, United States. *wluo@tamu.edu*

Abstract: Testing mediation effects has an important role in social and behavioral research. So far, the extant literature on multilevel mediation analysis has exclusively dealt with data with strictly nested structure. In reality, however, multilevel data may not always be strictly nested, but cross-classified (e.g., students cross-classified by schools and neighborhoods). This study will examine four common mediation designs in cross-classified multilevel data: (1) the 1à 1 à1 design, (2) the 2(A) à1à1 design, (3) the 2(A)à 2(A)à1 design, and (4) the 2(A)à 2(B)à1 design. The first three designs are direct extensions of their counterparts in strictly nested data. They can be analyzed using cross-classified Structural Equation Models. However, the fourth design is unique to cross-classified data because the initial cause is associated with one crossed factor, the mediator is associated with the other crossed factor, and the outcome is associated with level-1 units. Such design cannot be handled by cross-classified SEMs, therefore we propose a method that uses the Multiple Membership (MM) model and the Cross-classified Random Effects (CC) model to estimate and test the indirect effect in the 2(A)à 2(B) à1 design. The method will be illustrated using real data from the Early Childhood Longitudinal Study – Kindergarten Cohort (1998). Simulation studies will be conducted to examine the performance of the proposed method under various sample size conditions, effect sizes, and degrees of cross-classification.

Mon, June 27 (13:30-15:10) | TCP16

Recent Advances and Challenges in Analysis of Complex Biomedical Data

Organizer: Sijian Wang (University of Wisconsin-Madison)

Chair: Lingsong Zhang (Purdue University)

Latent Class Modeling using Matrix-valued Covariates with Application to Identifying Early Placebo Responders Based on EEG Signals

Bei Jiang

University of Alberta, Canada. *bei1@ualberta.ca*

Abstract: Latent class models are widely used to identify latent subgroups based upon one or more manifest variables. The probability of belonging to each subgroup can be simultaneously related to a set of measured covariates. In this paper, we extend existing latent class models to incorporate matrix covariates. This research is motivated by a placebo-controlled depression clinical trial. One study goal is to identify a subgroup of subjects who benefit from placebo effect (i.e., early placebo responders) as manifested by a clinical depression severity measure; and to relate the likelihood of belonging to this subgroup to baseline Electroencephalography (EEG) measurement that takes the form of a matrix. The proposed method is built upon a low rank Candecomp/Parafac (CP) decomposition to express the target coefficient matrix through low-dimensional latent variables, which effectively reduces the model dimensionality. We further adopt a Bayesian hierarchical modeling approach to estimating these latent variables, which provides a flexible way to incorporate prior knowledge on the patterns of covariate effect heterogeneity and provides a data-driven method of regularization. Simulation studies suggest that our proposed method is also robust against potentially misspecified rank in the CP decomposition. Finally in our motivating example, we show that the proposed method allows us to extract useful information from baseline EEG measurements that explains the likelihood of belonging to the early placebo responder subgroup.

Quantile Regression with Varying Coefficients for Functional Responses

Linglong Kong

University of Alberta, Canada. *lkong@ualberta.ca*

Abstract: With modern technology development, functional data are often observed in various scientific fields. Quantile regression has become an important statistical methodology. In this paper, we consider the estimation and inference about varying coefficients models for functional responses on quantile regression processes. We first propose to estimate the quantile smooth coefficient functions using local linear approximations, obtain the global uniform Bahadur representation of the estimator with respect to the time or the location and the quantile level, and show that the estimator converges weakly to a two-parameter continuous Gaussian process, and then we obtain asymptotic bias and mean integrated square error of smoothed individual functions and their uniform convergence rate under the given quantile level. We propose a global test for linear hypotheses of varying coefficient functions under quantile processes, and derive its asymptotic distribution under the null hypothesis; and also give their simultaneous confidence bands. For develop these inferences, some unknown error densities are estimated by the "residual-based" empirical distributions. A Monte Carlo simulation is conducted to examine the finite-sample performance of the proposed procedures. Finally, we illustrate the estimation and inference procedures of QRVC to diffusion tensor imaging data and ADHD-200 fMRI data. Joint work with Xingcai Zhou, Rohana Karunamuni, and Hongtu Zhu.

A Multi-scale Spatial Point Process Model for Stroke Lesion Segmentation on Multimodal MRI Data

Huiyan Sang

Texas A&M University, United States. *huiyan@stat.tamu.edu*

Abstract: Ischemic stroke is the third most frequent cause of death and a major cause of disability in industrial countries. In clinical practice, Diffusion weighted images (DWI), T1-weighted (T1W), T2-weighted (T2W) and fluid attenuated inversion recovery (FLAIR) images are often acquired to diagnose and monitor disease progression of strokes. However, manual segmentation of stroke lesions from these brain images is often a challenging and time consuming task and can only be performed by trained clinicians. In this paper, we propose an automated method to locate, segment and quantify stroke lesion areas using a new bias correction algorithm and a multi-scale 3D spatial point process clustering model.

We evaluate the performance of the proposed model using the Ischemic Stroke Lesion Segmentation Challenge data.

Integrative Analysis of High-dimensional Genomic Data

Sijian Wang

University of Wisconsin-Madison, United States. *wangs@stat.wisc.edu*

Abstract: Two types of integrations of multiple genomic datasets are considered in this talk. The first type of integration is based on the datasets from multiple studies. For example, an increasing amount of gene expression datasets for similar biomedical problems is available through public repositories. Integrating data from different but independent studies may facilitate discovery of new biological insights. We propose a meta-lasso method for gene selection with multiple expression data. Through a hierarchical decomposition on regression coefficients, our method not only borrows strength across multiple datasets to boost the power to identify important genes, but also keeps the selection flexibility among datasets to take into account data heterogeneity. Our method can incorporate exogenous information about gene function (such as pathway) into the modeling stage as well. The second type of integration is based on the datasets of multiple types. Recently, many genome-wide datasets capturing somatic mutation patterns, DNA copy number alterations, DNA methylation changes and gene expression are simultaneously obtained in the same biological samples. These samples render an integrated data resolution that may not be available with any single data type. We propose an iCluster+ method for pattern discovery (clustering) by integrating diverse data types. The core idea is motivated by the hypothesis that diverse molecular phenotypes can be predicted by a set of orthogonal latent variables that represent distinct molecular drivers, and thus can reveal tumor subgroups of biological and clinical importance.

Mon, June 27 (13:30-15:10) l TCP17

Recent Advances in Object Data Analysis

Organizer: Jie Peng (University of California, Davis)

Chair: Sungkyu Jung (University of Pittsburgh)

Sequential Change-point Detection Based on Nearest Neighbors

Hao Chen

University of California, Davis, United States. *hxchen@ucdavis.edu*

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.

Small Circle Distributions for Estimation of Rotational Axis from Directional Data

ByungWon Kim

University of Pittsburgh, United States. *byk4@pitt.edu*

Abstract: When shape changes or deformations of 3D objects occur, they can be represented by movements of directional vectors on the unit sphere. In particular, if major deformations can be assumed as rigid rotation, bending or twisting, the movements of vectors are concentrated on small circles on the unit sphere. Motivated by this physical observation, we propose two new probability distributions for multivariate directional data, aiming to model the small-circle concentrated data. These distributions need not to be rotationally symmetric, and can model dependencies among directions. We investigate statistical properties of the proposed distributions. Likelihood-based estimation algorithms and related large-sample tests for some important parameters are proposed. Our estimation procedures show better or comparable performances compared with other methods of small-circle fitting in a simulation study. The new distributions and their estimation procedures are demonstrated for 3D objects data, concerning skeletal representations of deformed ellipsoids and knee motions during gait.

Statistical Analysis of Trajectories on Riemannian Manifolds

Jingyong Su

Texas Tech University, United States. *jingyong.su@ttu.edu*

Abstract: In this research we propose to develop a comprehensive framework for registration and analysis of manifold-valued processes. Functional data analysis in Euclidean spaces has been explored extensively in literature. But we study a different problem in the sense that functions to be studied take values on nonlinear manifolds, rather than in vector spaces. Manifold-valued data appear frequently in shape and image analysis, computer vision, biomechanics and many others. If the data were contained in Euclidean space, one would use standard Euclidean techniques and there has been a vast literature on these topics. However, the non-linearity of the manifolds requires development of new methodologies suitable for analysis of manifold-valued data. We propose a comprehensive framework for joint registration and analysis of multiple manifold-valued processes. The goals are to take temporal variability into account, derive a rate-invariant metric and generate statistical summaries (sample mean, covariance etc.), which can be further used for registering and modeling multiple trajectories.

Multiscale Modeling of Hi-C Data

Rachel Wang

Stanford University, United States. *ryxwang@stanford.edu*

Abstract: Genome-wide chromosome conformation capture technology, broadly referred to as Hi-C, enables the generation of 3D genome contact maps and offers new insights into the spatial organization of genome. It is widely recognized that chromosomes form domains of enriched interactions playing significant roles in gene regulation and development. These domains include large-scale open or closed chromatin compartments and densely interacting, contiguous regions at sub-megabase scale known as topologically associating domains (TADs). Although a few algorithms have been proposed to detect TADs, developing statistical frameworks capable of incorporating the hierarchical nature of domains and discerning finer structures within TADs is still a nascent topic. We provide statistical analysis of Hi-C data to detect these multiscale domains and identify ones that are conserved across different cell lines. We compare our method with existing algorithms and show enrichment of key epigenetic markers at the domain boundaries.

Mon, June 27 (13:30-15:10) I CP01

Machine Learning Session 1

Chair: Xiaodan Fan (The Chinese University of Hong Kong)

Bayesian Bandwidth Selection in Nonparametric Models Based on Local Polynomial Regression

Zhongcheng Han

Southeast University, China. *hzcmaster@163.com*

Abstract: As we known, bandwidth selection is an very important step in nonparametric regression models. This paper presents a Bayesian approach to select bandwidth for nonparametric models based on local polynomial regression. We proved that the bandwidth selected by Bayesian approach is equivalent to CV's under the case of non-information prior. What is more, the bandwidth selection by Bayesian approach is not very sensitive to the prior information of the bandwidth. A MCMC simulation is executed to confirm our proposition, and then we applied the method to estimate price of the S&P 500 index options data.

Clustering Functional Data using Projection

Tung Pham

The University of Melbourne, Australia. *pham.t@unimelb.edu.au*

Abstract: We show that, in the functional data context, by appropriately exploiting the functional nature of the data, it is possible to cluster the observations asymptotically perfectly. We demonstrate that this level of performance can often be achieved by the k-means algorithm as long as the data are projected on a carefully chosen finite dimensional space. We propose an iterative algorithm to choose the projection functions in a way that optimises clustering performance, where, to avoid peculiar solutions, we use a weighted least-squares criterion. We apply our iterative clustering procedure on simulated and real data, where we show that it works well.

Optimal Estimation of Derivatives in Nonparametric Regression

Wenlin Dai

King Abdullah University of Science and Technology, Saudi Arabia. *wenlin.dai@kaust.edu.sa*

Abstract: We propose a simple framework for estimating derivatives without fitting the regression function in nonparametric regression. Unlike most existing methods that use the symmetric difference quotients, our method is constructed as a linear combination of observations. It is hence very flexible and applicable to both interior and boundary points, including most existing methods as special cases of ours. Within this framework, we define the variance-minimizing estimators for any order derivative of the regression function with a fixed bias-reduction level. For the equidistant design, we derive the asymptotic variance and bias of these estimators. We also show that our new method will, for the first time, achieve the asymptotically optimal convergence rate for difference-based estimators. Finally, we provide an effective criterion for selection of tuning parameters and demonstrate the usefulness of the proposed method through extensive simulation studies of the first- and second-order derivative estimators.

Modeling and Forecasting on-line Auction Prices: A Semi-parametric Regression Analysis

Weiwei Liu

Lanzhou University, China. *liuww@lzu.edu.cn*

Abstract: Interest in on-line auctions has been increasingly growing in recent years. There is an extensive literature on this topic, whereas modeling on-line auction price process constitutes one of the most active research areas. Most of the research, however, only focus on modeling price curves, ignoring the bidding process. In this paper, a semi-parametric regression model is proposed to model the on-line auction process. This model captures two main features of on-line auction data: changing arrival rates of bidding processes and changing dynamics of prices. An new inference procedure using B-splines is also established for parameters estimation. The proposed model is used to forecast the price of an on-line auction. The advantage of this proposed approach is that the price can be forecast dynamically and the prediction can be updated according to newly arriving information. The model is applied to the Xbox data with satisfactory forecasting property.

Sparse Regularization for Multiclass Functional Logistic Regression

Hidetoshi Matsui

Shiga University, Japan. *hmatsui@math.kyushu-u.ac.jp*

Abstract: Penalties with an L^1 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 L^1 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 decision boundaries. 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 are estimated by extending an existing algorithm, and then values of tuning parameters included in the regularization method is decided by a model selection criterion. We then apply the proposed method to the analysis of real data.

Mon, June 27 (15:30-17:10) l DL10 l Sponsor: Chinese Society of Probability and Statistics

Statistical Analysis for Social Network Data

Chair: Hoi Ying Wong (The Chinese University of Hong Kong)

Distinguished Lecturer: Hansheng Wang (Peking University)

Network Vector Autoregression

Hansheng Wang

Peking University, China. *hansheng@gsm.pku.edu.cn*

Abstract: We consider here a large-scale social network with a continuous response observed for each node at equally spaced time points. The responses from different nodes constitute an ultrahigh dimensional vector, whose time series dynamics is to be investigated. In the meanwhile, the network structure needs to be taken into consideration. To this end, we propose a network vector autoregressive (NAR) model. NAR models each node's response at a given time point as a linear combination of (a) its previous value, (b) the average of its connected neighbors, (c) a set of node-specific covariates, and (d) an independent noise. The corresponding coefficients are referred to as the momentum effect, the network effect, and the nodal effect respectively. The strictly stationary solutions of NAR model are obtained for fixed and diverging network size respectively. In order to estimate the NAR model, an ordinary least squares type estimator is developed, and its asymptotic properties are investigated. We further illustrate the usefulness of the NAR model through a number of interesting potential applications. Simulation studies and an empirical example are presented to demonstrate the performance of the newly proposed methodology.

Least Squares Estimation of Spatial Autoregressive Models for Large-scale Social Networks

Danyang Huang

Renmin University of China, China. *dyhuang89@126.com*

Abstract: Due to the rapid development of various online social network websites, the spatial autoregressive (SAR) model is becoming an important tool in social network analysis. However, two major bottlenecks remain in analyzing large-scale networks. First, existing methods are computationally infeasible for very large networks (e.g. Facebook has over 700 million active users); second, there is a lack of a good network sampling scheme to ensure valid estimation and inference for the entire network structure based on the observed data only. To address both of these challenges, we propose a novel least squares estimator (LSE) approach. For networks with sparse network structures, the computational complexity of the LSE is shown to be linear in network size, making it scalable to huge networks. In theory, the LSE is square root of n consistent and asymptotically normal under certain regularity conditions. Furthermore, the sampling scheme derived from the LSE can automatically adjust autocorrelation between sampled and unsampled units, ensuring consistent estimation and valid statistical inference of the SAR model. Numerical results based on simulated and real data are presented as well.

Multivariate Spatial Autoregression for Large Scale Social Network

Xuening Zhu

Peking University, China. *xueningzhu@pku.edu.cn*

Abstract: The rapid growth of social network platforms generates a large amount of social network data. As a result, multivariate responses and the corresponding predictors can be collected for social network users. To statistically model such type of data, the multivariate spatial autoregression (MSAR) model is proposed and studied. To estimate the model, the maximum likelihood estimator (MLE) is obtained under certain technical conditions. However, it is found that the computational cost of MLE is expensive. In order to fix this problem, a least squares type estimator is developed. The corresponding asymptotic properties are investigated. To gauge the finite sample performance of the proposed estimators, a number of numerical studies are conducted. Lastly, a Sina Weibo dataset is analyzed for illustration purpose.

Mon, June 27 (15:30-17:10) | DL17 | Sponsor: IMS

Sparse Learning from Covariance

Chair: Cun-Hui Zhang (Rutgers University)

Distinguished Lecturer: Jianqing Fan (Princeton University)

Robust Statistical Learning from Covariance

Jianqing Fan

Princeton University, United States. *jqfan@princeton.edu*

Abstract: This talks will introduce some recent developments on the robust estimation of large covariance matrices. These include methods based on factor models, elliptical models, RA-quadratic, and projected PCA. These methods are then applied to robust sparse regression, deep principal learning, and sparse regression with highly dependent covariates. Their theoretical properties will also be unveiled.

Discussion of the Distinguished Lecture

Harrison Zhou

Yale University, United States. *huibin.zhou@yale.edu*

Abstract: In this talk, we will discuss Professor Jianqing Fan's Distinguished Lecture.

Large-scale Mean-variance Portfolio Optimization

Xinghua Zheng

The Hong Kong University of Science and Technology, Hong Kong. *xhzheng@ust.hk*

Abstract: In this talk we discuss some recent progress on large-scale mean-variance portfolio optimization. We illustrate how the major tools in high-dimensional statistics including sparse regression, random matrix theory, factor models etc. can be used to solve this important problem.

Based on joint works with Mengmeng Ao and Yingying Li.

Mon, June 27 (15:30-17:10) l IP07 l Sponsor: IMS

Big Data Integration

Organizer: Jason Fine (The University of North Carolina at Chapel Hill)

Chair: Jason Fine (The University of North Carolina at Chapel Hill)

Integrative Statistical Analysis and Exploration of Mixed-type Massive Data

Susan Wilson

The University of New South Wales and The Australian National University, Australia. *sue.wilson@anu.edu.au*

Abstract: Massive data are becoming ubiquitous in a large number of fields, including health care and biomedical research. To fully utilise such data, new methods are needed. One fundamental challenge is how to integrate and explore big data that may be of different types. For example, the data may have an extremely large number of components consisting of a huge number of (i) continuous, and (ii) categorical variables. One popular paradigm arises in modern omics data analysis where there are measures (continuous) on many thousands of gene expression loci as well as many thousands of (categorical) genomic variants (such as SNPS, single nucleotide polymorphisms) across the whole genome, and there may be further, other measurements (say clinical). An information theoretic measure of association to deal with integrative analysis of such data has been developed. It is relatively robust compared with common measures of association, and has been incorporated into an R package that is fast and easily parallelised, and proving popular. Further, an approximate permutation style test has been advanced to control the overall type I error rate. The final output can be used for downstream exploration, such as network visualisations.

This is joint research with Drs Chris Pardy and Paul Lin.

Shared Informative Factor Models for Integration of Multi-platform Bioinformatic Data

Jianhua Hu

The University of Texas MD Anderson Cancer Center, United States. *jhu@mdanderson.org*

Abstract: High-dimensional omic data derived from different technological platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms and to determine personalized health treatments. Numerous studies have integrated multi-platform omic data; however, few have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations. We propose a statistical framework of shared informative factor models that can jointly analyze multi-platform omic data and explore their associations with a disease phenotype. The common disease-associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype. Extensive simulation studies demonstrate the performance of the proposed method in terms of biomarker detection accuracy via comparisons with three popular regularized regression methods.

Fusion Learning in Data Integration

Peter XK Song

University of Michigan, United States. *pxsong@umich.edu*

Abstract: Combining datasets collected from similar studies becomes increasingly popular in biomedical research to achieve larger sample sizes and greater statistical power. One of major methodological challenges in such data integration pertains to data heterogeneity that complicates statistical analysis in many ways. The traditional meta-analysis is based on a critical assumption of common parameters across studies, which is often violated in practice. Another approach based on the mixed-effects model assumes that all inter-study parameters are different, leading to possible over-parametrization. In this talk I will focus on regression analysis of combined data of smaller datasets in which subject-level information is available. We develop a method of fusion learning that enables simultaneously the identification of homogeneous parameter clusters, without using the method of hypothesis testing, and the estimation of the fused parameters. Utilizing solution paths and tuning parameters, we propose a new graphic tool, named as fusogram, to visualize the progression of parameter fusion as well as to quantify the friction of fusion in connection to data heterogeneity. All related theory is established in the framework of estimating functions, so generalized linear models for cross-sectional data and generalized estimating equations for longitudinal data are included in the development. Both simulation studies and real world data examples are used to motivate and illustrate the proposed methodology.

Optimal Estimation for Quantile Regression with Functional Response

Xiao Wang

Purdue University, United States. *wangxiao@purdue.edu*

Abstract: Quantile regression with functional response and scalar covariates has become an important statistical tool for many neuroimaging studies. In this paper, we study optimal estimation of varying coefficient functions in the framework of reproducing kernel Hilbert space. Minimax rates of convergence under both fixed and random designs are established. We have developed easily implementable estimators which are shown to be rate-optimal. Simulations and real data analysis are conducted to examine the finite-sample performance. This is a joint work with Zhengwu Zhang, Linglong Kong, and Hongtu Zhu.

Mon, June 27 (15:30-17:10) l IP12 l Sponsor: ICSA

Longitudinal Data Analysis

Organizer: Jialiang Li (National University of Singapore)

Chair: Jialiang Li (National University of Singapore)

On Last Observation Carried Forward and Asynchronous Longitudinal Regression Analysis

Hongyuan Cao

University of Missouri, United States. *caohong@missouri.edu*

Abstract: In many longitudinal studies, the covariates and response are often intermittently observed at irregular, mismatched and subject-specific times. Last observation carried forward (LOCF) is one of the most commonly used methods to deal with such data when covariates and response are observed asynchronously. However, this can lead to considerable bias. In this paper, we propose a weighted LOCF estimation using asynchronous longitudinal data for the generalized linear model. We further generalize this approach to utilize previously observed covariates in addition to the most recent observation. In comparison to earlier methods, the current methods are valid under weaker assumptions on the covariate process and allow informative observation times which may depend on response even conditional on covariates. Extensive simulation studies provide numerical support for the theoretical findings. Data from an HIV study is used to illustrate our methodology.

Variable Selection for Fixed and Random Effects in Generalized Linear Mixed Models

Liming Xiang

Nanyang Technological University, Singapore. *LMXiang@ntu.edu.sg*

Abstract: Generalized linear mixed models (GLMMs) are widely used for analysis of longitudinal or clustered data due to their ability of directly accounting for intracluster dependence and modeling different types of data. In the presence of many covariates, variable selection becomes important. Typical existing works consider selection on either fixed or random effects separately. To avoid possible model misspecification, we propose a selection procedure for both fixed and random effects based on approximate inference in GLMMs. We select and estimate fixed effects by means of iteratively reweighted penalized profile likelihood, in the meantime select and estimate random effects through reweighted penalized restricted posterior likelihood. We show the oracle property of the proposed procedure and demonstrate its performance via simulation studies and real data examples. This is a joint work with Tonghui Yu.

Efficient Estimation in Semivarying Coefficient Models for Longitudinal / Clustered Data

Toshio Honda

Hitotsubashi University, Japan. *t.honda@r.hit-u.ac.jp*

Abstract: In semivarying coefficient modeling of longitudinal/clustered data, 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 using the true within-subject covariance matrices, which are often unavailable in reality. Thus we propose an estimator when the covariance matrices are unknown and depend only on the index variable with some specification assumption. Then we estimate the covariance matrices using residuals obtained from a preliminary estimation based on working independence and both spline and local linear regression. Then, using the covariance matrix estimates, we employ spline regression again to obtain our final estimator. It achieves the semiparametric efficiency bound under normality assumption and has the smallest asymptotic covariance matrix among a class of estimators even when normality is violated. Our theoretical results hold either when the number of within-subject observations diverges or when it is uniformly bounded. We also considered nonparametric component estimation. The proposed method is compared with the working independence estimator and some existing method via simulations and application to a real data example. This is joint work with Ming-Yen Cheng and Jialiang Li.

Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories with Application to Cocaine Abuse Treatment Data

Yehua Li

Iowa State University, United States. *yehuali@iastate.edu*

Abstract: In a cocaine dependence treatment study, we have paired binary longitudinal trajectories that record the cocaine use patterns of each patient before and after a treatment. To better understand the drug-using behaviors among the patients, we propose a general framework based on functional data analysis to jointly model and cluster these paired non-Gaussian longitudinal trajectories. Our approach assumes that the response variables follow distributions from the exponential family, with the canonical parameters determined by some latent Gaussian processes. To reduce the dimensionality of the latent processes, we express them by a truncated Karhunen-Lóeve (KL) expansion allowing the mean and covariance functions to be different across clusters. We further represent the mean and eigenfunctions functions by flexible spline bases, and determine the orders of the truncated KL expansions using data-driven methods. By treating the cluster membership as a missing value, we cluster the cocaine use trajectories by a likelihood-based approach. The cluster membership and parameter estimates are jointly estimated by a Monte Carlo EM algorithm with Gibbs sampling steps. We discover subgroups of patients with distinct behaviors in terms of overall probability to use, binge verses periodic use pattern, etc. The joint modeling approach also sheds new lights on relating relapse behavior to baseline pattern in each subgroup.

Mon, June 27 (15:30-17:10) l IP15 l Sponsor: IISA

Limit Theorems in Stochastic Analysis

Organizer: Arnab Ganguly (Louisiana State University)

Chair: Amarjit Budhiraja (The University of North Carolina at Chapel Hill)

Moderate Deviation Principles for Stochastic Differential Equations

Arnab Ganguly

Louisiana State University, United States. *aganguly@lsu.edu*

Abstract: Moderate and large deviation principles involve estimating the probabilities of rare events. In particular, they often help to assess the quality of approximating models obtained through law of large number-type results. The talk will focus on a weak convergence based approach to moderate deviation principles for stochastic differential equations with jumps.

On a Rescaling Transformation for Stochastic Partial Differential Equations

Michael Roeckner

Bielefeld University, Germany. *roeckner@math.uni-bielefeld.de*

Abstract: In the talk we shall give a survey on the rescaling transformation to stochastic partial differential equations. We shall also present some recent new applications.

Large Deviations for Multi-scale Jump-diffusions

Rohini Kumar

Wayne State University, United States. *rkumar@math.wayne.edu*

Abstract: We obtain large deviation results for a two time-scale model of jump-diffusion processes. The processes on the two time scales are fully inter-dependent, the slow process has small perturbative noise and the fast process is ergodic. We will discuss both probabilistic and PDE methods that can be used to obtain these results. Some applications of these results to Finance will also be discussed.

This is joint work with Jean-Pierre Fouque, Jin Feng, Martin Forde and Lea Popovic.

Dissipation and High Disorder

Kunwoo Kim

Pohang University of Science and Technology, South Korea. *kkim@math.utah.edu*

Abstract: In this talk, we consider a limiting behavior of the total mass of a semi-discrete stochastic heat equation on sZ^d. We show that almost surely the total mass goes to 0 under high disorder whereas it is strictly positive under weak disorder. This is based on joint work with Le Chen, Michael Cranston, and Davar Khoshnevisan.

Mon, June 27 (15:30-17:10) | IP20 | Sponsor: Korea

Nonparametric Inference in Econometric Models

Organizer: Kyusang Yu (Konkuk University)

Chair: Kyusang Yu (Konkuk University)

Operational Time and In-sample Density Forecasting

Young K. Lee

Kangwon National University, South Korea. *youngklee@kangwon.ac.kr*

Abstract: We consider a new structural model for in-sample density forecasting. In-sample density forecasting is to estimate a density function on a region where the density is observed and then re-use the estimated density while estimating the density on a region it is not observed. Our structural assumption is that the density is a product of one-dimensional functions with one function sitting on the scale of a transformed space of observations. The transformation involves another unknown one-dimensional function, so that our model is formulated via a known smooth function of three underlying unknown one-dimensional functions. We present an innovative way of estimating the one-dimensional functions and show that all the estimators of the three components achieve the optimal one-dimensional rate of convergence. We illustrate how one can use our approach by analysing a real data from the insurance business. We also investigate the finite sample performance of the method via a simulation study.

A New Estimator for Stochastic Frontier Models

Hohsuk Noh

Sookmyung Women's University, South Korea. *hsnoh@sookmyung.ac.kr*

Abstract: We consider the estimation of a nonparametric stochastic frontier model with composite error density which is known up to a finite parametric vector. In this setting, Martins-Filho and Yao (2015) proposed a profile-likelihood estimator and showed its efficiency. However, the estimator is very hard to implement in practice since the profiled nonparametric component cannot be expressed as a closed form of the parametric component. Hence, we develop an estimator which is as efficient as their estimator but has easy implementation.

Partial Identification in Regression Discontinuity Designs with Manipulated Running Variables

Christoph Rothe

Columbia University, United States. *cr2690@columbia.edu*

Abstract: The regression discontinuity (RD) design has become a popular empirical strategy in economics. The distinct feature of RD is that treatment assignment is fully determined by whether the value of a continuous running variable lies below or above a known cutoff. This provides a transparent way to identify and estimate treatment effects for individuals at the cutoff under relatively mild assumptions. The idea behind this is local randomization: it is essentially random whether a unit falls just below or just above the cutoff. The key assumption underlying local randomization is that units cannot manipulate the value of their running variable in order to guarantee or avoid assignment to treatment. If such manipulation was possible, it would be unlikely that units just below and just above the cutoff would be comparable because of self-selection. Thus manipulation is likely to break down local randomization and consequently the identification of treatment effects. In this paper, we propose an approach to obtain sharp (that is, best-possible) bounds on average treatment effects under manipulation. We also propose new and non-standard methods to estimate these bounds in practice, and show how to conduct inference. We apply the approach to study the effect of unemployment insurance on unemployment duration in Brazil.

A Semiparametric Intraday GARCH-X Model

Melanie Schienle

Karlsruhe Institute of Technology, Germany. *melanie.schienle@kit.edu*

Abstract: We propose a multiplicative component model for intraday volatility. The model consists of a semiparametric and parametric component. The former captures the well-documented intraday seasonality of volatility as well as the impact of the state of the limit order book, utilizing a semiparametric additive structure. The parametric component accounts for short-run fluctuations by means of a unit GARCH specification. The model is estimated by a simple and easy to implement algorithm, consisting of smooth-backfitting and QML steps. Further, we provide an automatic data-driven procedure for bandwidth choice. We show the asymptotic properties of the estimator and document its finite sample performance in a comprehensive simulation study. Finally, our empirical application based on high-frequency data for NASDAQ equities investigates non-linearities in the relationship between the limit order book and subsequent return volatility and underlines the usefulness of including order book variables for out-of-sample forecasting performance.

Mon, June 27 (15:30-17:10) l IP57

Recent Advances and Challenges of Big Data Inference with Complex Structures

Organizer: Dan Yang (Rutgers University)

Chair: Dan Yang (Rutgers University)

Linear Regression Analysis in a Model Free Setting

Linda Zhao

University of Pennsylvania, United States. *lzhao@wharton.upenn.edu*

Abstract: Linear regression analysis is often applied to populations that do not satisfy the conventional assumptions of linearity, homoscedasticity and normality of residuals. Furthermore, the classical theory considers the covariates as fixed constants, whereas in most applications they are actually random variables. The basic perspectives thus involve assumption-lean statistical populations that include random covariates that may have an unspecified distribution, and then applying assumption-rich statistical models as approximations. Results concerning both inference about the linear regression coefficients and about best linear predictive inference will be covered. This is joint work with a group of people at Penn.

The Power Prior: Theory and Applications

Ming-Hui Chen

University of Connecticut, United States. *ming-hui.chen@uconn.edu*

Abstract: The power prior has been widely used in many applications covering a large number of disciplines. The power prior is intended to be an informative prior constructed from historical data. It has been used in clinical trials, genetics, health care, psychology, environmental health, engineering, economics, and business. It has also been applied for a wide variety of models and settings, both in the experimental design and analysis contexts. We review its theoretical properties, variations in its formulation, statistical contexts for which it has been used, applications, and its advantages over other informative priors. We review models for which it has been used, including generalized linear models, survival models, and random effects models. Statistical areas where the power prior has been used include model selection, experimental design, hierarchical modeling, and conjugate priors. Frequentist properties of power priors in posterior inference are established and a simulation study is conducted to further examine the empirical performance of the posterior estimates with power priors. Real data analyses are given illustrating the power prior.

Two-sample Tests for High-dimensional Linear Regression with an Application to Gene and Environment Interactions

Yin Xia

The University of North Carolina at Chapel Hill, United States. *xiayin@email.unc.edu*

Abstract: Motivated by important applications in genomics, we consider in this talk global and coordinatewise tests for the comparisons of two high-dimensional linear regression models. A procedure for testing the equality of the two regression vectors globally is proposed and shown to be particularly powerful against sparse alternatives. In addition, we introduce a multiple testing procedure for identifying unequal coordinates while controlling the false discovery rate (FDR) and false discovery proportion (FDP). Theoretical justifications are provided to guarantee the validity of the proposed tests and optimality results are established under sparsity assumptions on the regression coefficients. Simulation results show that the proposed tests maintain the desired error rates under the null and have good power under the alternative at moderate sample sizes. The procedures are applied to the Framingham Offspring study to investigate the interactions between smoking and cardiovascular related genetic mutations important for an inflammation marker.

Nested Nonnegative Cone Analysis

Lingsong Zhang

Purdue University, United States. *lingsong78@icloud.com*

Abstract: Motivated by the analysis of nonnegative data objects, a novel Nested Nonnegative Cone Analysis (NNCA) approach is proposed to overcome some drawbacks of existing methods. The application of traditional PCA/SVD method to nonnegative data often cause the approximation matrix leave the nonnegative cone, which leads to non-interpretable and sometimes nonsensical results. The nonnegative matrix factorization (NMF) approach overcomes this issue, however the NMF approximation matrices suffer several drawbacks: (1) the factorization may not be unique, (2) the resulting approximation matrix at a specific rank may not be unique, and (3) the subspaces spanned by the approximation matrices at different ranks may not be nested. These drawbacks will cause troubles in determining the number of components and in multi-scale (in ranks) interpretability. The NNCA approach proposed in this paper naturally generates a nested structure, and is shown to be unique at each rank. Simulations are used in this paper to illustrate the drawbacks of the traditional methods, and the usefulness of the NNCA method.

Mon, June 27 (15:30-17:10) | TCP12

Analysis of Reliability and Survival Data

Organizer: Tony Hon Keung Ng (Southern Methodist University)

Chair: Tony Hon Keung Ng (Southern Methodist University)

Lifetime Inference for Highly Reliable Products Based on Skew-normal Accelerated Destructive Degradation Test Model

Chien-Tai Lin

Tamkang University, Taiwan. *chien@mail.tku.edu.tw*

Abstract: The accelerated destructive degradation test (ADDT) method provides an effective way to assess the reliability information of highly reliable products whose quality characteristics degrade over time, and can be taken only once on each tested unit during the measurement process. Conventionally, engineers assume that the measurement error follows the normal distribution. However, degradation models based on this normality assumption often do not apply in practical applications. To relax the normality assumption, the skew-normal distribution is adopted in this study because it preserves the advantages of the normal distribution with the additional benefit of flexibility with regard to skewness and kurtosis. Here, motivated by polymer data, we propose a skew-normal nonlinear ADDT model, and derive the analytical expressions for the product's lifetime distribution along with its corresponding th percentile. Then, the polymer data are used to illustrate the advantages gained by the proposed model. Finally, we addressed analytically the effects of model misspecification when the skewness of measurement error are mistakenly treated, and the obtained results reveal that the impact from the skewness parameter on the accuracy and precision of the prediction of the lifetimes of products is quite significant.

Proportional Hazards Model for Accelerated Life Testing Data from One-shot Devices

Man Ho Ling

The Hong Kong Institute of Education, Hong Kong. *amhling@ied.edu.hk*

Abstract: For devices with long lifetimes, accelerated life-tests are commonly used to induce quick failures. A link function relating stress levels and lifetime is then applied to extrapolate the lifetimes of units from accelerated conditions to normal operating conditions. Because data from one-shot devices do not contain any lifetimes, a standard reliability analysis with a parametric distributional assumption on lifetimes may be sensitive to violations of the model assumption. For this reason, we consider a proportional hazards model to analyze one-shot device testing data collected from constant-stress accelerated life-tests. A Monte Carlo simulation study shows that the proposed flexible semi-parametric model provides a good insight into the estimation of reliability under normal (typical) operating conditions.

Inference for the Generalized Pareto Distribution and its Application

Hideki Nagatsuka

Chuo University, Japan. *hideki@indsys.chuo-u.ac.jp*

Abstract: The generalized Pareto distribution (GPD) is widely used to model exceedances over thresholds. The estimation of the parameters of the GPD is a difficult problem and existing methods for estimating parameters have theoretical or computational defects. In this talk, we will introduce an approach to inference for the GPD, proposed by Nagatsuka and Balakrishnan (2015), which can remedy such difficulty and will present some applications based on real data sets.

The EM Algorithm for One-shot Device Testing with Competing Risk under Different Lifetimes Distributions

Hon Yiu So

McMaster University, Canada. *sohy@math.mcmaster.ca*

Abstract: In this talk, we extend the recent works of Balakrishnan and Ling by introducing a competing risk model into a one-shot device testing analysis under accelerated life test setting. Expectation maximization (EM) algorithms are developed for the estimation of model parameters under different lifetime distributions. Extensive Monte Carlo simulations are carried out to assess the performance of the proposed method of estimation. The advantages of the EM algorithms over the traditional Fisher scoring method are displayed through simulation.

Strategic Allocation of Test Units in an Accelerated Degradation Test Plan

Zhisheng Ye

National University of Singapore, Singapore. *yez@nus.edu.sg*

Abstract: Degradation is often defined in terms of the change of a key performance characteristic over time. It is common to see that the initial performance of the test units varies and it is strongly correlated with the degradation rate. Motivated by a real application in the semiconductor sensor industry, this study advocates an allocation strategy in accelerated degradation test (ADT) planning by capitalizing on the correlation information. In the proposed strategy, the initial degradation levels of the test units are measured and the measurements are ranked. The ranking information is used to allocate the test units to different factor levels of the accelerating variable. More specifically, we may prefer to allocate units with lower degradation rates to a higher factor level in order to hasten the degradation process The allocation strategy is first demonstrated using a cumulative-exposure degradation model. Likelihood inference for the model is developed. The optimum test plan is obtained by minimizing the large sample variance of a lifetime quantile at nominal use conditions. Various compromise plans are discussed. A comparison of the results with those from traditional ADTs with random allocation reveals the value of the proposed allocation rule. To demonstrate the broad applicability, we further apply the allocation strategy to two more degradation models which are variants of the cumulative-exposure model.

Mon, June 27 (15:30-17:10) l TCP13

Theory and Applications of Tensor Variate Data Analysis

Organizer: Toshio Sakata (Kyushu University)

Chair: Kohei Adachi (Osaka University)

Inference for Tensor Elliptical Distributions

Mohammad Arashi

Shahrood University of Technology, Iran. *m_arashi_stat@yahoo.com*

Abstract: The multilinear normal distribution is a widely used tool in tensor analysis of magnetic resonance imaging (MRI). Diffusion tensor MRI provides a statistical estimate of a symmetric 2nd-order diffusion tensor, for each voxel within an imaging volume. In this talk, the tensor matrix elliptical (TME) distribution is introduced as an extension to the multilinear normal (MLN) distribution. Some properties including the characteristic function and distribution of affine transformations are given. An integral representation connecting densities of TME and MLN distributions is exhibited that is used in deriving the expectation of any measurable function of a TME variate. Some inferential issues are also addressed.

Universal Subspaces for Local Unitary Groups of Fermionic Systems

Lin Chen

Beihang University, China. *linchen@buaa.edu.cn*

Abstract: Motivated by the development of quantum information in recent years, in this talk we study the universal subspace of fermionic (i.e., antisymmetric) space in terms of the local unitary (LU) group, i.e., every vector in the space can be converted into a vector of the subspace under the group. The space is spanned by the Slater determinant vectors, just like that the Hilbert space is spanned by tensor product vectors. So they have similar problems in tensor analysis. Below is the concrete description of our talk. We show that the single occupancy space is the universal subspace of tripartite antisymmetric subspace. When the local dimension is even, we further prove that the universal subspace has minimum dimension and find out the explicit choice of Slater determinants for the subspace. For 4-partite antisymmetric subspace, we prove that the single occupancy subspace is not universal. We construct the explicit example, namely the well-known BCS states that are not equivalent to any single occupancy states.

Sparse Three-way PCA for Selecting the Optimal Model Between Tucker2 and Parafac

Hiroki Ikemoto

Osaka University, Japan. *theresawill1112@gmail.com*

Abstract: Three-way principal component analysis (3WPCA) procedures have been used in various fields, such as psychometrics and chemometrics, for analyzing a three-way data array of objects × variables × sources. In 3WPCA, a three-way data array is approximately decomposed into component matrices and a core array. A core array describes the relationships between components in different component matrices. Popular 3WPCA models are the Tucker2 model, the Tucker3 model and the Parafac model. Among them, the least restrictive is the Tucker2 model and the most restrictive is the Parafac model. In the Tucker2 model, a core array is unconstrained. On the other hand, the core array in Parafac is strongly constrained in that the core slices in the array are forced to be diagonal matrices. These models have an advantage and a drawback. The Tucker2 model fits data well in the least squares sense and a core array tends to be complicated for interpretation. Conversely, the constrained core array in Parafac is superior in interpretability and its fitness to data is likely to be worse. In this study, we propose a new procedure by which we can find a suitably constrained intermediate model between Tucker2 and Parafac. In the proposed procedure, the Tucker2 loss function is minimized subject to a specified number of core elements being exact zero, while their locations are unknown. Therefore, the optimal locations of zero elements and nonzero parameter values are simultaneously estimated. We name the proposed procedure sparse core Tucker2 (ScTucker2), as the matrices with a number of zeros are said to be sparse. We present a procedure for selecting a suitable number of zero elements as well as an alternating least squares algorithm for ScTucker2. The effectiveness of ScTucker2 are demonstrated via a simulation study and a real data example. This is a joint work with Henk A.L. Kiers and Kohei Adachi.

One-sided Tests for Tensor Variate Normal Distributions

Manabu Iwasa

Kumamoto University, Japan. *iwasa@gpo.kumamoto-u.ac.jp*

Abstract: The problem of one-sided test of the mean vector of a normal distribution has been studied by several researchers. First, the problem was investigated when the covariance structure is known. Pioneering works had been done by Bartholomew (1959) and Kudo (1963). Later, Perlman (1969) considered the problem for the multivariate normal distribution when the covariance matrix is a completely unknown positive definite matrix. He found the fact that the more powerful test can be constructed by weakening the restriction in alternative hypothesis. Sasabuchi (2003, 2007) considered one-sided tests for the mean vectors of several multivariate normal distributions and revealed the similar phenomena. Recently, the author developed the result to matrix variate normal distributions. In this article, we extend them to tensor variate normal distributions.

Typical Ranks of Tensors Over the Real Number Field and Determinantal Ideals

Mitsuhiro Miyazaki

Kyoto University of Education, Japan. *g53448@kyokyo-u.ac.jp*

Abstract: High dimensional array data, that is, tensors are widely studied in various fields of pure and applied mathematics from various points of view. The rank of a tensor is an invariant which measures the complexity of the tensor and is a generalization of the rank of a matrix, a 2-dimensional tensor. Consider the set of m by n matrices, where m and n are positive integers such that m is less than or equal to n. Then almost all m by n matrices have rank m. Thus one may think an m by n matrix whose rank is not m an exceptional one. The phenomenon for 3 or more dimensional tensors is quite different and if one works over the real number field, there may be multiple non-exceptional ranks, which are called typical ranks. We characterize whether n by p by m tensors have multiple typical ranks when p is large relative to m and n by corresponding to an n by p by m tensor a determinantal ideal of a certain mn-p by n matrix in a polynomial ring over the real number field.

Mon, June 27 (15:30-17:10) | TCP14

Recent Developments and Applications of Structural Equation Modeling (SEM) Techniques I: Measurement Invariance and Latent Growth Models

Organizer: Oi-Man Kwok (Texas A&M University)

Chair: Oi-Man Kwok (Texas A&M University)

Model Specification Search for Identifying the Optimal Growth Trajectory in Latent Growth Models

MinJung Kim

The University of Alabama, United States. *mjkim@ua.edu*

Abstract: We conducted a Monte Carlo simulation to investigate the optimal strategy to search for the true mean trajectory under the latent growth modeling (LGM) framework. In this study, the effectiveness of different starting models was examined in terms of the mean and within-subject variance-covariance (V-C) structure model. The results showed that specifying the most complex (i.e., unstructured) within-subject V-C structure with the use of LRT, ΔAIC, and ΔBIC achieved the highest recovery rate (>85%) of the true mean trajectory. Implications of the findings and limitations will be discussed.

Multilevel Factorial Invariance in Ordered Categorical Measures

Ehri Ryu

Boston College, United States. *ehri.ryu.1@bc.edu*

Abstract: This study will describe a general procedure for testing factorial invariance in multilevel confirmatory factor analysis model with ordered categorical measures. Factorial invariance concerns whether the measures have equivalent measurement properties in distinctive groups. In multilevel research, the distinctive group membership may exist at the lower level or at the higher level in the multilevel structure. The lower-level group membership introduces additional complexities because there is dependency in the data between distinctive groups. This study will use multilevel mixture model framework proposed by Asparouhov and Muthén (2012) to solve this problem. The performance of test statistics will be empirically evaluated in a simulation study.

Testing Longitudinal Measurement Invariance using Majority Votes through a Sequential Procedure

Jiun-Yu Wu

National Chiao Tung University, Taiwan. *jiunyu.rms@gmail.com*

Abstract: Longitudinal measurement invariance (MI) implies that a measure assesses similar constructs in the same way at different ages or assessment waves. However, measurement invariance across age is often assumed but rarely tested in applied studies. In this study, we presented a sequential procedure to test longitudinal MI by comparing the model for each subsequent year with that of the previous year(s). Satorra-Bentler scaled differential chi-square test ( ) was conducted to take the data dependency nature into consideration. To evaluate the quality of longitudinal measurement invariance, four criteria of invariance analysis were used including changes in CFIs (∆CFI; Cheung & Rensvold, 2002), TLIs (∆TLI; Little, 1997), Root Deterioration per Restriction Index (RDR; Browne & Du Toit, 1992), and the Expected Cross-Validation Index difference (Browne & Cudeck, 1993; Oort, 2009). The MI decision is reached when the majority of criteria are within the suggested thresholds as evidence of measurement invariance. A nine annual dataset of Teacher Network Relationship Inventory (TNRI) was used to demonstrate longitudinal stability of the measurement structure with the proposed procedure. As a result of applying this procedure, the TNRI was found to meet configural, metric and scalar invariance assumptions in longitudinal setting within CFA framework.

Testing Factorial Invariance with Severely Unbalanced Samples

Myeongsun Yoon

Texas A&M University, United States. *myoon@tamu.edu*

Abstract: When factorial invariance is examined, large imbalances in group sizes can affect the results of the study because the chi-square statistics include a weighting by sample sizes. In consequence, the groups with larger sample size will have more weight in determining the final solution and the impact of violations of invariance in the smaller groups might be reduced. The implication of the result is that violations of invariance might not be easily detected if sample sizes of the two groups are severely unbalanced. The present study examined effect of sample size differences in groups on power in detecting violations of factorial invariance using simulated data sets. Also, we proposed a bootstrapping sample method to address severely unbalanced sample size issue in factorial invariance studies and compared the proposed approach with a random selection method that has frequently been used.

Mon, June 27 (15:30-17:10) I CP02

Machine Learning Session 2

Chair: Tung Pham (The University of Melbourne)

Comparing Two Nonparametric Regression Curves in the Presence of Long Memory in Covariates and Errors

Fang Li

Indiana University-Purdue University Indianapolis, United States. *fali@iupui.edu*

Abstract: This paper discusses the problem of testing the equality of two nonparametric regression functions against two-sided alternatives when both the common covariate and the two error processes form long memory moving averages. The proposed test is based on a marked empirical process of the differences between the response variables. We discuss the asymptotic null distribution of this process and the consistency of the test for a class of general alternatives. We also conduct a Monte Carlo simulation to study the finite sample level and power behavior of the test at some alternatives.

Robust Regression for Highly Corrupted Response by Shifting Outliers

Yoonsuh Jung

The University of Waikato, New Zealand. *yoonsuh@waikato.ac.nz*

Abstract: Outlying observations are often disregarded at the sacrifice of degrees of freedom or downsized via robust loss functions (for example, Huber's loss) to reduce the undesirable impact on data analysis. In this paper, we treat the outlying status of each observation as a parameter and propose a penalization method to automatically adjust the outliers. The proposed method shifts the outliers towards the fitted values, while preserve the non-outlying observations. We also develop a generally applicable algorithm in the iterative fashion to estimate model parameters and demonstrate the connection with the maximum likelihood based estimation procedure in the case of least squares estimation. We establish asymptotic property of the resulting parameter estimators under the condition that the proportion of outliers do not vanish as sample size increases. We apply the proposed outlier adjustment method to ordinary least squares and lasso-type of penalization procedure and demonstrate its empirical value via numeric studies. Furthermore, we study applicability of the proposed method to two robust estimators, Huber's robust estimator and Huberized lasso, and demonstrate its noticeable improvement of model fit in the presence of extremely large outliers.

Schwarz-type Model Comparison for LAQ Models

Shoichi Eguchi

Kyushu University, Japan. *s-eguchi@math.kyushu-u.ac.jp*

Abstract: The classical Bayesian information criterion (BIC) is derived through the stochastic expansion of marginal likelihood function under suitable regularity condition. However, despite of its popularity, mathematical validity of BIC for possibly misspecified models with complicated dependence structure is often ignored. Thus it is important to extend the reach of the classical BIC with rigorous theoretical foundation. In this talk, we will give a general result about the stochastic expansion of the logarithmic marginal quasi-likelihood associated with a class of locally asymptotically quadratic (LAQ) family of statistical experiments. Based on the expansion, we propose the quasi-Bayesian information criterion (QBIC), which prevails even when the corresponding M-estimator is of multi-scaling type and the asymptotic quasi-information matrix is random, as well as when statistical model is misspecified. Importantly and especially, this extension allows us to consider a wide range of non-ergodic stochastic-process models observed at high-frequency. This study is a joint work with Hiroki Masuda.

Kernel Entropy Estimation for Linear Processes

Hailin Sang

The University of Mississippi, United States. *sang@olemiss.edu*

Abstract: Let $X_n = \sum_{i=0}^{\infty} a_i \varepsilon_{n-i}$, where the ε_i are i.i.d. centered random variables taking vales in R^d and $\sum_{i=0}^{\infty} |a_i| < \infty$. Assume that f is the probability density function of X_n. We consider the estimation of the quadratic functional $\int f^2(x)dx$. It. It is shown that, under certain conditions, the estimator

$$\frac{2}{n(n-1)h_n^d} \sum_{1 \le i < j \le n} K\left(\frac{X_i - X_j}{h_n}\right)$$

is asymptotically efficient. For i.i.d. case, Giné and Nickl (Bernoulli, 2008) applied a convolution method to obtain the bias, and used the Hoeffding's decomposition for *U*-statistics to study the stochastic part. Instead, for linear processes, we apply the Fourier transform on the kernel function to derive the asymptotic properties. The result and the method have application to L_2^2 divergence between distributions of two linear processes.

Day 2
Tue, June 28

Tue, June 28 (08:30-10:10) l DL13 l Sponsor: Local Host

Recent Development of Statistical Inference Under Dimension Reduction Structure

Chair: Xin Yuan Song (The Chinese University of Hong Kong)

Distinguished Lecturer: Lixing Zhu (Hong Kong Baptist University)

Adaptive-to-model Test for Parametric Single-index Models: A Dimension Reduction Approach

Lixing Zhu

Hong Kong Baptist University, Hong Kong. *lzhu@hkbu.edu.hk*

Abstract: Local smoothing testing based on multivariate nonparametric regression estimation is one of the main model checking methodologies in the literature. However, the relevant tests suffer from typical curse of dimensionality, resulting in slow convergence rates to their limits under the null hypothesis and less deviation from the null hypothesis under alternative hypotheses. This problem prevents tests from maintaining the significance level well and makes tests less sensitive to alternative hypotheses. In this paper, a model-adaption concept in lack-of-fit testing is introduced and a dimension-reduction model-adaptive test procedure is proposed for parametric single-index models. The test behaves like a local smoothing test, as if the model were univariate. It is consistent against any global alternative hypothesis and can detect local alternative hypotheses distinct from the null hypothesis at a fast rate that existing local smoothing tests can achieve only when the model is univariate. Simulations are conducted to examine the performance of our methodology. An analysis of real data is shown for illustration. The method can be readily extended to global smoothing methodology and other testing problems.

Sufficient Dimension Reduction for Longitudinal Data

Annie Peiyong Qu

University of Illinois at Urbana-Champaign, United States. *anniequ@illinois.edu*

Abstract: Correlation structure contains important information about longitudinal data. Existing sufficient dimension reduction approaches assuming independence may lead to substantial loss of efficiency. We apply the quadratic inference function to incorporate the correlation information and apply the transformation method to recover the central subspace. The proposed estimators are shown to be consistent and more efficient than the ones assuming independence. In addition, the estimated central subspace is also efficient when the correlation information is taken into account. We compare the proposed method with other dimension reduction approaches through simulation studies, and apply this new approach to longitudinal data for an environmental health study.

A Post-screening Diagnostic Study in Sufficient Dimension Reduction for Ultrahigh Dimensional Data

Liping Zhu

Renmin University of China, China. *zhuliping.stat@yahoo.com*

Abstract: Sufficient dimension reduction is a paradigm that combines the idea of linear dimension reduction with the concept of sufficiency. In this article we propose a consistent lack-of-fit test to examine whether or not replacing the original ultrahigh dimensional covariates with a given number of linear combinations will result in loss of regression information. To attenuate spurious correlations which are often seen in ultrahigh dimensional covariates and may substantially inflate type-I error rates, we suggest to randomly split the observations into two halves. In the first halve of observations we screen out as many irrelevant covariates as possible. This helps us reduce the ultrahigh dimensionality to a moderate scale. In the second halve we perform a lack-of-fit test for conditional independence within the context of sufficient dimension reduction. This data-splitting strategy helps us retain the type-I error rate pretty well. We propose a new statistic to test conditional independence, and show that our propose test procedure is *n*-consistent under the null and root-*n*-consistent under the alternative hypothesis. Our proposed test procedure is consistent in the sense that it has nontrivial power against all feasible alternatives. In addition, we suggest a bootstrap procedure to decide critical values and show that our bootstrap procedure is consistent. We demonstrate the effectiveness of our test procedure through comprehensive simulations and an application to the rats red-eye data set.

Tue, June 28 (08:30-10:10) l IP60 l Sponsor: IMS

Special Session in Memory of Peter Hall

Organizer: Alan Welsh (Australian National University)

Chair: Alan Welsh (Australian National University)

Invited Speakers: Susan Wilson, The University of New South Wales and The Australian National University
Matt Wand, University of Technology Sydney
Jianqing Fan, Princeton University
Byeong Park, Seoul National University
Bingyi Jing, The Hong Kong University of Science and Technology
Ming-Yen Cheng, National Taiwan University

Tue, June 28 (08:30-10:10) I IP16 I Sponsor: Bernoulli Society

Models and Inference for Complex Extreme Events

Organizer: Anthony Davison (Ecole Polytechnique Fédérale de Lausanne)

Chair: Richard A. Davis (Columbia University)

Generalized Additive Modeling of Nonstationary Multivariate Extremes

Miguel de Carvalho

Pontificia Universidad Católica de Chile, Chile. *mdecarvalho@mat.puc.cl*

Abstract: In this talk I will introduce smooth nonstationary generalized additive modeling for nonstationary extremal dependence structures. I start by arguing that time-varying spectral measures are a natural concept for modeling dynamic structures of dependence between extremes of random variables. Our approach is constructed from time-varying versions of well-known parametric models for the spectral density of a multivariate extreme value distribution, such as the logistic, Dirichlet, and the Husler–Reiss model. Fitting is conducted by a maximum penalized approximated likelihood estimator based on an approximation to the multivariate extreme value density earlier proposed by Cooley, Davis, and Naveau (2007). The methods are applied to simulated data, and a case study in Finance is used to motivate the need for theory and methods, as well as to illustrate the main concepts introduced along the talk. Joint work with Valérie Chavez-Demoulin and Linda Mhalla.

References

Cooley, D., Davis, R., and Naveau, P. (2007), "Prediction for Max-Stable Processes via an Approximated Conditional Density," Technical Report.

Exact Simulation of Max-stable Processes

Sebastian Engelke

Ecole Polytechnique Fédérale de Lausanne, Switzerland. *sebastian.engelke@epfl.ch*

Abstract: Max-stable processes play an important role as models for spatial extreme events. Their complex structure as the pointwise maximum over an infinite number of random functions makes simulation highly nontrivial. Algorithms based on finite approximations that are used in practice are often not exact and computationally inefficient. We will present two algorithms for exact simulation of a max-stable process at a finite number of locations. The first algorithm generalizes the approach by Dieker and Mikosch for Brown-Resnick processes and it is based on simulation from the spectral measure. The second algorithm relies on the idea to simulate only the extremal functions, that is, those functions in the construction of a max-stable process that effectively contribute to the pointwise maximum. We study the complexity of both algorithms and prove that the second procedure is always more efficient. Moreover, we provide closed expressions for their implementation that cover the most popular models for max-stable processes and extreme value copulas. For simulation on dense grids, an adaptive design of the second algorithm is proposed.

Full Likelihood Inference For Max-stable Distributions Based on a Stochastic EM Algorithm

Raphael Huser

King Abdullah University of Science and Technology, Saudi Arabia. *raphael.huser@kaust.edu.sa*

Abstract: Max-stable distributions are widely used for the modeling of multivariate extreme events, as they arise as natural limits of renormalized componentwise maxima of random vectors. However, when the dimension is large, the number of terms involved in the likelihood function becomes extremely large, making it intractable for classical inference. In practice, composite likelihoods are often used instead, but suffer from a loss in efficiency. In this talk, an alternative approach to perform full likelihood inference based on an EM algorithm is explored, where an additional random partition associated to the occurrence times of maxima is introduced. Treating this partition as a missing observation, the completed likelihood becomes simple and a (stochastic) EM algorithm may be used to maximize the full likelihood. The performance of this novel approach will be illustrated with numerical results based on the logistic model. If time allows, the performance will also be assessed for more complex models.

Joint work with Clement Dombry, Marc Genton and Mathieu Ribatet.

Probabilities of Concurrent Extremes

Stilian Stoev

University of Michigan, United States. *sstoev@umich.edu*

Abstract: Suppose that one measures precipitation at several synoptic stations over multiple days. We say that extremes are concurrent if the maximum precipitation over time is achieved simultaneously over all stations, e.g., on a single day. This is a strong indication of spatial dependence where a single "storm event" causes the extremes. Under general conditions, the finite sample concurrence probability converges to an asymptotic quantity, deemed extremal concurrence probability. This was first established in Hashorva and Hüsler (2005) (see also Stephenson and Tawn (2005) and Wadsworth and Tawn (2015)).

Using Palm calculus, we establish general expressions for the extremal concurrence probability through the max-stable process emerging in the limit of the componentwise maxima of the sample. Explicit forms of the extremal concurrence probabilities are obtained for various max-stable models and several estimators are introduced. In particular, we prove that the pairwise extremal concurrence probability for max-stable vectors is precisely equal to the Kendall's tau. The estimators are evaluated by using simulations and applied to study the expected area of concurrence regions of temperature extremes in the United States. The results demonstrate that concurrence probability can provide a powerful new perspective and tools for the analysis of the spatial structure and impact of extremes.

Joint work with Clement Dombry and Mathieu Ribatet.

References

Dombry, Cl., Ribatett, M. and Stoev, S. (2014). Probabilities of Concurrent Extremes. Preprint.

Hashorva, E. and J. Hüsler (2005). Multiple maxima in multivariate samples. Statist. Probab. Lett. 75(1), 11–17.

Stephenson, A. and J. Tawn (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. Biometrika 92(1), 213–227.

Tue, June 28 (08:30-10:10) l IP28 l Sponsor: Local Host

Asymptotic Theory

Organizer: Qi-Man Shao (The Chinese University of Hong Kong)

Chair: Qi-Man Shao (The Chinese University of Hong Kong)

Parisi Formula, Disorder Chaos and Fluctuation for the Ground State Energy in the Spherical Mixed p-spin Models

Wei-Kuo Chen

University of Minnesota, United States. *mathzygmund@gmail.com*

Abstract: Spin glasses are disordered spin systems originated from the desire of understanding the strange magnetic behaviors of certain alloys in physics. As mathematical objects, they are often cited as examples of complex systems and have provided several fascinating structures and conjectures. In this talk, we will focus on the spherical mixed p-spin mean-field spin glass model. We will present the Parisi formula and some fluctuation properties for the maximum energy. In addition, we will discuss results concerning the chaotic nature of the location of the maximum energy under small perturbations to the disorder. This talk is based on a joint work with Arnab Sen.

Testing Independence with High-dimensional Correlated Samples

Weidong Liu

Shanghai Jiao Tong University, China. *weidongl@sjtu.edu.cn*

Abstract: Testing independence among a number of (ultra) high-dimensional random samples is a fundamental and challenging problem. By arranging n identically distributed p-dimensional random vectors into a p x n data matrix, we investigate the testing problem on independence among columns under the matrix-variate normal modeling of the data. We propose a computationally simple and tuning free test statistic, characterize its limiting null distribution, analyze the statistical power and prove its minimax optimality. As an important by-product of the test statistic, a ratio-consistent estimator for the quadratic functional of covariance matrix from correlated samples is developed. We further study the effect of correlation among samples to an important high-dimensional inference problem - large-scale multiple testing of Pearson's correlation coefficients. It can be shown that blindly using classical inference result based on the sample independence assumption will lead to many false discoveries, which suggests the need for conducting independence testing before applying existing methods. To address the challenge arising from the correlation among samples in correlation test, we propose a "sandwich estimator" of Pearson's correlation coefficient by de-correlating the samples, based on which the resulting multiple testing procedure asymptotically controls the overall false discovery rate at the nominal level while maintaining good statistical power. Both simulated and real data experiments are carried out to demonstrate the advantages of the proposed methods. (joint with Xi Chen)

Berry-Esseen Bound for Exchangeable Pairs

Zhuosong Zhang

The Chinese University of Hong Kong, Hong Kong. *zhuosongzhang@foxmail.com*

Abstract: A new Berry-Esseen bound for exchangeable pairs is established without the boundedness assumption of $W-W'$. Optimal convergence rate for normal and non-normal approximation can be achieved using this result. Our main result can be applied in many fields of models, such as quadratic forms, simple random sampling, general Curie-Weiss model, mean field Heisenberg model, colored graph model and limit theorem for prime factors.

Spurious Discoveries in High Dimension

Wenxin Zhou

Princeton University, United States. *zhouwenxin1986@gmail.com*

Abstract: Many data-mining and statistical machine learning algorithms have been developed to select a subset of covariates to associate with a response variable. Spurious discoveries can easily arise in high-dimensional data analysis due to enormous possibilities of such selections. How can we know statistically our discoveries better than those by chance? In this paper, we define a measure of goodness of spurious fit, which shows how good a response variable can be fitted by an optimally selected subset of covariates under the null model. It coincides with the maximum spurious correlation for linear models and can be regarded as a generalized maximum spurious correlation. We derive the asymptotic distribution of such goodness of spurious fit for generalized linear models and L^1-regression. Such an asymptotic distribution depends on the sample size, ambient dimension, the number of variables used in the fit, and the covariance information. It can be consistently estimated by multiplier bootstrapping and used as a benchmark to guard against spurious discoveries.

Tue, June 28 (08:30-10:10) l IP32 l Sponsor: Japan

Distributions on Convex Cones and Sets of Permutations

Organizer: Satoshi Kuriki (The Institute of Statistical Mathematics)

Chair: Donald Richards (The Pennsylvania State University)

Multivariate Distribution Theory by Holonomic Gradient Method

Akimichi Takemura

Shiga University, Japan. *takemura@stat.t.u-tokyo.ac.jp*

Abstract: The holonomic gradient method proposed in Nakayama et al (2011) utilizes partial differential equations for computing parameterized integrals and very useful in studying the multivariate distribution theory. In this talk, I report recent results of applications of holonomic gradient method to multivariate distribution theory.

Generation of Random Permutations: Algorithms, Implementation and Probabilistic Analysis

Hsien-Kuei Hwang

Academia Sinica, Taiwan. *hkhwang@stat.sinica.edu.tw*

Abstract: Random permutations are indispensable in diverse areas of science and engineering. We present a few simple and efficient algorithms for generating random permutations of large number of elements, say 10^9. These algorithms have all been proposed in the statistical literature but most of them have remained practically unknown. We propose a complete probabilistic analysis of their cost in terms of bit complexity, as well as more modern implementations of them (including the extension to a parallel computing environment and benchmarks).

Some Distributions Associated with the Cone of Positive Semidefinite Matrices and their Applications

Satoshi Kurik

The Institute of Statistical Mathematics, Japan. *kuriki@ism.ac.jp*

Abstract: Let A be a standard Gaussian random matrix in the space Sym(n) of n x n symmetric (or Hermitian) matrices. Let PD(n) be the cone of positive semidefinite matrices in Sym(n). In this talk, we derive the distribution of the squared distance between the random matrix A and the cone PD(n). This distribution appears as the null distribution of the likelihood ratio criterion for testing multivariate variance components. In real and complex normal population cases, the distributions are mixtures of chi-square distributions with weights expressed in terms of the Pfaffian and the determinant, respectively. Moreover, when the size n of the matrix goes to infinity, by modifying Johansson's (1998) central limit theorem for eigenvalues of random matrices, the limiting distribution is proved to be Gaussian. This property of limiting Gaussianity was conjectured in previous literature (e.g., Amemiya, Anderson and Lewis, 1990).

Joint work with Tomoyuki Shirai and Trinh Khanh Duy of Kyushu University

Totally Positive Exponential Families

Caroline Uhler

Massachusetts Institute of Technology, United States. *caroline.uhler@ist.ac.at*

Abstract: We discuss properties of distributions that are multivariate totally positive of order two (MTP2). In particular, we prove that any independence model generated by an MTP2 distribution is a compositional semigraphoid which is upward-stable and singleton-transitive. As a consequence of this result, we obtain that the MTP2 constraints (which are convex constraints) imply a Markov structure, and hence sparsity, without the need of a tuning parameter. We discuss this interesting alternative for modeling in the high-dimensional setting in the framework of MTP2 exponential families.

Tue, June 28 (08:30-10:10) | IP35 | Sponsor: India

Survey Sampling and Related Methodologies

Organizer: Sanjay Chaudhuri (National University of Singapore)

Chair: Malay Ghosh (University of Florida)

An Empirical Likelihood Based Estimator for Respondent Driven Sampled Data

Sanjay Chaudhuri

National University of Singapore, Singapore. *sanjay@stat.nus.edu.sg*

Abstract: We discuss an empirical likelihood based estimator of population means applicable to data obtained from a respondent driven sampling procedure. Our estimator directly uses the second order weights of selection and constructs a composite empirical likelihood to estimate the parameter of interest. This estimate is asymptotically unbiased and normally distributed. Analytic expression of the asymptotic standard errors can be obtained which can also be estimated from the data using a sandwich estimator. Using real life social network data, we show that, our estimator produces confidence intervals with far better coverages than the existing estimators. This is a joint work with Mark Handcock from University of California, Los Angeles.

Small Area Model Selection

Singdhansu Chatterjee

University of Minnesota, United States. *chatterjee@stat.umn.edu*

Abstract: Complex statistical models that have multiple sources of dependencies and variability in the observations are of primary importance in studying data from multiple disciplines. These include spatial, temporal, spatio-temporal, various mixed effects and other statistical models. Of special importance among such models are those that are useful for studying problems where there is limited directly observed data, for example, as in small area models. In this talk we present a new resampling-based method that can be used for simultaneous variable selection and inference in several complex models, including small area and other mixed eff ects models. Theoretical results justifying the proposed resampling schemes will be presented, followed by simulations and real data examples. This talk is based on research involving several students and collaborators from multiple institutions.

Multiple Imputation and/or Calibration in Two-phase Designs?

Thomas Lumley

The University of Auckland, New Zealand. *t.lumley@auckland.ac.nz*

Abstract: Two-phase epidemiologic studies, where a new exposure is measured on a designed subsample of an existing cohort, are traditionally analysed by survey methods. Calibration of weights allows information from the whole cohort to be used in estimation. Multiple imputation, invented as a technique for missing data, has also been suggested for filling in the incomplete data on the new exposure. I will describe how these are related, show that multiple imputation can be used with calibration to obtain the (asymptotically) optimal calibration estimator, and discuss the efficiency: robustness tradeoffs in choosing between multiple imputation and calibration.

Multiple Imputation using the Weighted Finite Population Bayesian Bootstrap

Michael Elliott

University of Michigan, United States. *mrelliot@umich.edu*

Abstract: Multiple imputation (MI) is a principled method to deal with item-level missing data that has become increasingly popular in the public health and social science investigations where data production is often based on complex sample surveys. However, existing software packages and procedures typically do not incorporate complex sample design features in the imputation process. Failure to account for design features, particularly sampling weights, can introduce bias on final estimates. Recent work to accommodate complex sample designs in imputation includes the sample design in the formulation of the imputation model, which typically requires strong model assumptions and be difficult to implement in practice. We propose a new method to incorporate complex sample designs in MI, using recent work (Dong et al. 2014) that extends finite population Bayesian bootstrap to generate synthetic populations from a posterior predictive distribution in a fashion that inverts the complex sampling design features (weights, clusters, and stratification) and generates simple random samples from a superpopulation point of view, making adjustment on the complex data so that they can be analyzed as simple random samples. We then perform conventional parametric MI for missing data at the second step using readily available imputation software designed for an SRS sample. A new combining rule for the point and variance estimates is derived to make valid inferences based on the two-step procedure. We evaluate the performance of the new method in comparison with the fully model-based method through simulation, and provide several example applications. Results show that the new method is more robust to model misspecification and generally yields lower RMSE than the fully model-based method.

Tue, June 28 (08:30-10:10) | TCP03

Statistics and Computing for Complex Dependent Systems

Organizer: Kengo Kamatani (Osaka University)

Chair: Teppei Ogihara (The Institute of Statistical Mathematics)

Robust Estimation for Sparse Gaussian Graphical Modeling

Kei Hirose

Osaka University, Japan. *hirose@sigmath.es.osaka-u.ac.jp*

Abstract: Gaussian graphical modeling has been widely used to explore various network structures, such as gene regulatory networks and social networks. We often use a penalized maximum likelihood approach with the L^1 penalty for learning a high-dimensional graphical model. However, the penalized maximum likelihood procedure is sensitive to outliers. To overcome this problem, we introduce a robust estimation procedure based on the γ-divergence. The parameter estimation procedure is constructed using the Majorize-Minimization algorithm, which guarantees that the objective function monotonically decreases at each iteration. Simulation studies show that our procedure performs much better than the existing methods, in particular, when the contamination rate is large.

Multilevel Sequential Monte Carlo Samplers

Ajay Jasra

National University of Singapore, Singapore. *staja@nus.edu.sg*

Abstract: In this talk we consider the approximation of expectations w.r.t. probability distributions associated to the solution of partial differential equations (PDEs); this scenario appears routinely in Bayesian inverse problems. In practice, one often has to solve the associated PDE numerically, using, for instance finite element methods and leading to a discretisation bias, with the step-size level h_L. In addition, the expectation cannot be computed analytically and one often resorts to Monte Carlo methods. In the context of this problem, it is known that the introduction of the multilevel Monte Carlo (MLMC) method can reduce the amount of computational effort to estimate expectations, for a given level of error. This is achieved via a telescoping identity associated to a Monte Carlo approximation of a sequence of probability distributions with discretization levels $\infty>h_0>h_1\cdots>h_L$. In many practical problems of interest, one cannot achieve an i.i.d. sampling of the associated sequence of probability distributions. A sequential Monte Carlo (SMC) version of the MLMC method is introduced to deal with this problem. It is shown that under appropriate assumptions, the attractive property of a reduction of the amount of computational effort to estimate expectations, for a given level of error, can be maintained within the SMC context. The approach is numerically illustrated on a Bayesian inverse problem.

On Asymptotics of Multivariate Non-Gaussian Quasi-likelihood

Hiroki Masuda

Kyushu University, Japan. *hiroki@math.kyushu-u.ac.jp*

Abstract: We consider (semi-)parametric inference for a class of stochastic differential equation (SDE) driven by a locally stable Levy process, focusing on multivariate setting and some computational aspects. The process is supposed to be observed at high frequency over a fixed time domain. This setting naturally gives rise to a theoretically fascinating quasi-likelihood which brings about a novel unified estimation strategy for targeting a broad spectrum of driving Levy processes. The limit experiment is mixed normal with a clean-cut random information structure, based on which it is straightforward to make several conventional asymptotic statistical decisions. The infill-asymptotics adopted here makes the popular Gaussian quasi-likelihood useless, while instead enabling us not only to incorporate any exogenous and/or observable endogenous data into the trend and/or scale coefficients without any difficulty, but also to sidestep most crucial assumptions concerning the long-term stability such as ergodicity and moment boundedness. The proposed quasi-likelihood estimator is known to be asymptotically efficient in some special cases.

Some Properties of the Mixed Preconditioned Crank-Nicolson Algorithm

Kengo Kamatani

Osaka University, Japan. *kamatani@sigmath.es.osaka-u.ac.jp*

Abstract: We study some interesting properties of a general purpose MCMC, the mixed preconditioned Crank-Nicolson (MpCN) algorithm.

High-dimensional asymptotics, reversibility of the proposal kernel, random-walk property and ergodicity are studied.

We compare MpCN and other MCMC algorithms by these properties and provide some information for the choice of these MCMC algorithms.

Tue, June 28 (08:30-10:10) | TCP08

Stochastic Processes and Related Topics

Organizer: Dongsheng Wu (The University of Alabama in Huntsville)

Chair: Dongsheng Wu (The University of Alabama in Huntsville)

Small Value Probabilities for Supercritical Multitype Branching Processes

Weijuan Chu

Nanjing University, China. *chuwj@nju.edu.cn*

Abstract: Morters P. and Ortgiese M. (2008) presented a probabilistic method to obtain the small value probability of supercritical Galton-Watson process. The new method is intuitive and heuristic, thus convenient to be extended to other models. In this paper, we try to generalize it to multi-type branching processes to prove its small value probability, which was obtained in O.D. Jones (2004) by considering the property of generating function of the offspring.

Some Aspects of the Rosenblatt Sheet

Guangjun Shen

Anhui Normal University, China. *gjshen@163.com*

Abstract: In this talk, we study the problem of the approximation in law of the Rosenblatt sheet. We prove the convergence in law of four families of process to the Rosenblatt sheet: the first one is martingale differences, the second one constructed from a Poisson process in the plane and the third one is the random walks, the last one is based on the multiple Wiener integrals. This is a joint work with Litan Yan, Xiuwei Yin, Qian Yu, Dongjin Zhu.

Stepwise Estimation of Ergodic Levy Driven Stochastic Differential Equation

Yuma Uehara

Kyushu University, Japan. *ma214003@math.kyushu-u.ac.jp*

Abstract: Recent development of data-handling technique allows us to obtain high-frequency data from various time-varying phenomena. Especially biological, technological and financial field there are a lot of such phenomena whose driving noise exhibits non-Gaussian behavior. To incorporate it in statistical modeling, Levy driven stochastic differential equation (SDE) plays an important role. However any unified estimation strategy has not been established as yet because of the diversity of Levy process.

In this talk we consider high-frequency samples from a multi-dimensional ergodic Levy driven SDE model. We assume that the coefficients is known up to a finite-dimensional parameter and the driving Levy noise has finite moments of any order. Our goal is to estimate unknown parameters in coefficients and a functional parameter contained in Levy measure corresponding to the driving Levy noise. For this purpose we estimate these targets separately: first estimate scale parameter by drift-free Gaussian quasi-likelihood function; second by using estimator of scale parameter we estimate drift parameter; finally we construct the Euler residual building on Euler-Maruyama approximation and estimate a functional parameter.

We will present the standard asymptotic normality of our estimators with correction matrix constructed only by observed data and moment convergence of parameters in coefficients which is crucial, among others, for the derivation of some information criteria. Our multistage estimation procedure effectively reduces the high-computational load of optimization and permits overlap of parameter components. Further, this procedure only requires the moment existence of noise distribution, hence it has advantage in terms of the robustness of noise misspecification and is valid for the skewed noise driven Levy-driven SDE.

This is a joint work with Hiroki Masuda.

Quantile Regression Process: Nonparametric and Partially Linear Asymptotics

Shih-Kang Chao

Purdue University, United States. *skchao74@purdue.edu*

Abstract: Quantile regression process (QRP) reveals complete information for the whole conditional distribution. When combined with Kolmogorov-Smirnov or Cramer-von-Mises statistics, QRP can be applied to construct tests for treatment effect or conditional stochastic dominance. We consider the weak convergence for QRP arisen from three popular quantile models: general series approximation model (including parametric linear model using covariates with diverging dimensionality), nonparametric spline model and partial linear model. Weak convergence is also studied for important functionals of QRPs: the quantile function, the rearrangement operator and the derivative of the quantile function. New Bahadur representations are derived and may be interesting in their own right. This is a joint work with Guang Cheng from Purdue University and Stanislav Volgushev from Cornell University.

Tue, June 28 (08:30-10:10) l TCP23

Recent Advances in Time-to-Event Analysis

Organizer: Tony Sit (The Chinese University of Hong Kong)

Chair: Chi Wing George Chu (The Chinese University of Hong Kong)

Pseudo Value Method for Ultra High-dimensional Semiparametric Models with Life-time Data

Tony Sit

The Chinese University of Hong Kong, Hong Kong. *tony.sit@cuhk.edu.hk*

Abstract: Technology advances facilitate collection of high-dimensional covariate information including microarray, proteomic and SNP data. Challenges have frequently been encountered when scientists attempt to understand the association, if any, between the high-dimensional covariates of the subjects and the survival time, which is complicated due to censoring. In this work, we develop a new rank-based approach that enables us to tackle the variable selection problem for various semiparametric models for life-time data via the novel pseudo value method. While there has only been sure independence screening (SIS) for ultra-high dimensional data modelled by Cox proportional hazards models in literature, our methodology can handle a much broader class of semiparametric models including general transformation models and the accelerated failure time model. Numerical studies have demonstrated promising performance that is comparable to the (iterative) sure independence screening (SIS). Our method was also applied to analyse Diffuse large-B-cell lymphoma data, which discovered potential genes that can be influential.

Efficient Estimation and Inference of Quantile Regression Under Biased Sampling

Gongjun Xu

University of Minnesota, United States. *xuxxx360@umn.edu*

Abstract: Biased sampling occurs frequently in economics, epidemiology and medical studies either by design or due to data collecting mechanism. Failing to take into account the sampling bias usually leads to incorrect inference. We propose a unified estimation procedure and a computationally fast resampling method to make statistical inference for quantile regression with survival data under general biased sampling schemes, including but not limited to the length-biased sampling, the case-cohort design and variants thereof. We establish the uniform consistency and weak convergence of the proposed estimator as a process of the quantile level. We also investigate more efficient estimation using the generalized method of moments and derive the asymptotic normality. We further propose a new resampling method for inference, which differs from alternative procedures in that it does not require to repeatedly solve estimating equations. It is proved that the resampling method consistently estimates the asymptotic covariance matrix. The unified framework proposed in this paper provides researchers and practitioners a convenient tool for analyzing data collected from various designs. Simulation studies and applications to real data sets are presented for illustration.

Confidence Intervals for High-dimensional Cox Models

Yi Yu

University of Cambridge, United Kingdom. *y.yu@statslab.cam.ac.uk*

Abstract: The purpose of this paper is to construct confidence intervals for high-dimensional Cox proportional hazards regression models where the number of time-dependent covariates can be larger than the sample size. The definition of the one-step estimator is similar to those in van de Geer et al. (2014) and Zhang and Zhang (2014), but since in the Cox regression model, the Hessian matrix is based on time-dependent covariates in censored risk sets, the technical difficulties are fundamentally different. We present the related theoretical results, algorithms and extensive numerical experiments in this paper. This is joint work with Jelena Bardic (UCSD) and Richard Samworth (Cambridge).

End-point Sampling

Wen Yu

Fudan University, China. *yuwen820523@163.com*

Abstract: Retrospective sampling designs, including case-cohort and case-control designs, are commonly used for failure time data in the presence of censoring. In this paper, we propose a new retrospective sampling design, called end-point sampling, which improves the efficiency of the case-cohort and case-control designs. The regression analysis is conducted using the Cox model. Under different assumptions, the maximum likelihood approach with the computational aid from the EM algorithm, as well as the inverse probability weighting approach are developed respectively to estimate the regression parameters. The resulting estimators are proved to be consistent and asymptotically normal. Simulation and real data studies show favorable evidence for the proposed design in comparison with the existing ones.

Tue, June 28 (08:30-10:10) I TCP06

Geometric Aspects of Statistical Methods

Organizer: Tomonari Sei (The University of Tokyo)

Chair: Tomonari Sei (The University of Tokyo)

A Decision Theoretic Property of Conditional Normalized Maximum Likelihood Distribution

Yoshihiro Hirose

The University of Tokyo, Japan. *hirose@mist.i.u-tokyo.ac.jp*

Abstract: We consider distribution prediction, where we observe a data and estimate the distribution of a future data based on the observation. The estimate of the distribution is called a prediction distribution. Our target is Conditional Normalized Maximum Likelihood (CNML) distribution. CNML is a generalization of Normalized Maximum Likelihood (NML) distribution. NML is the minimax distribution with respect to the regret. The regret is the difference between a candidate distribution and the best distribution for a virtually-observed value. Similarly, CNML is the minimax distribution with respect to a conditional regret. Grunwald (2007) introduced three versions of CNMLs corresponding to three types of conditional regrets. Based on an observed data, a conditional regret compares a prediction distribution with the best distribution for a virtually-observed value. In this talk, we are interested in whether CNMLs are admissible or not under some criterion. The admissibility is a basic concern in statistical decision theory. We are also interested in the minimaxity of CNMLs. CNML is originally minimax with respect to the conditional regret. However, it is not clear that it is also minimax under other criterion, e.g., the Kullback-Leibler divergence and conditional regret risks. The result depends on statistical models we assume. From the geometrical view point, a statistical model forms a manifold. We try to describe properties of CNMLs in terms of the geometry.

Information Geometry of Anomalous Statistics

Hiroshi Matsuzoe

Nagoya Institute of Technology, Japan. *matsuzoe@nitech.ac.jp*

Abstract: Anomalous statistics is a statistics for deformed exponential families. A deformed exponential family is a generalization of exponential family, and it was introduced in the theory of complex systems. In the case of exponential family, random variables and parameters are embedded into a probability distribution using an exponential function. In the deformed case, such objects are embedded using a suitable monotone function, called a deformed exponential function. Though deformed exponential functions and deformed logarithm functions play important roles in anomalous statistics, these functions do not satisfy the law of exponents in general. From this reason, deformed algebraic structures are introduced, and one can find that notions of the standard expectations and the standard independences are not natural for anomalous statistics. As a consequence, these notions are naturally modified due to the deformed algebraic structures.

In this talk, after summarizing preliminary facts of deformed exponential families, we discuss deformed expectations and independences of random variables. In information geometry, a Fisher metric and alpha-connections are known as the standard Riemannian metric and affine connections for statistical models, respectively. However, these are not suitable for anomalous statistics. Hence we construct information geometric structures on a deformed exponential family from the viewpoint of unbiasedness of estimating functions.

Symmetries in Experimental Design and Linear Estimators

Kentaro Tanaka

Seikei University, Japan. *tanaken@econ.seikei.ac.jp*

Abstract: In this paper, we provide a machine learning method to obtain an optimal design and show that we can classify the types of the linear estimators generated by the design by using Groebner bases. First, we show that the problem of constructing an optimal design matrix can be transformed into a problem of the group lasso. The problem here is that the optimal design obtained in that way strongly depends on the choice of the tuning parameters in the group lasso. This means that we need to choose tuning parameters appropriately to obtain the sparse solution such as orthogonal arrays as the solution of the problem of the group lasso. In order to overcome this problem, we need to classify the types of the linear estimators generated by the design. We explain that the method of Groebner bases of the toric ideal given by the design matrix is useful to characterize the linear estimators. Finally, we show that the types of the linear estimators generate a homological structure. We also provide an application of this method to high-dimensional data.

On Submanifolds of Textile Set

Ushio Tanaka

Osaka Prefecture University, Japan. *utanaka@mi.s.osakafu-u.ac.jp*

Abstract: The textile set is defined from the textile plot proposed by Kumasaka and Shibata (2008), which is a powerful tool for visualizing high dimensional data. The textile plot is based on a parallel coordinate plot, where the ordering, locations and scales of each axis are simultaneously chosen so that all connecting lines, each of which signifies an observation, are aligned as horizontally as possible. The plot transforms a data matrix in order to delineate a parallel coordinate plot.

The aim of this study is to investigate the structure of the textile set from a geometrical point of view. The set is a disjoint union of classifications, which are naturally defined from the transformed data. As one of results on such geometrical characterization of the classifications, we will show that each individual classification becomes a regular submanifold of the textile set under certain conditions. Then, from the differential geometrical viewpoint, it is natural to observe the relation between curvature and topology of the classification. To this end, it could be expected that a geodesic of the classification plays a significant role.

The present study is companion to our study: Geometric Properties of Textile Plot (2015).

Tue, June 28 (08:30-10:10) l CP14

Data Order Structure Session 1

Chair: Chun Yip Yau (The Chinese University of Hong Kong)

Quasi Hidden Markov Model and its Applications in Change-point Problems

Zhengxiao Wu

Singapore Management University, Singapore. *zxwu@smu.edu.sg*

Abstract: In a Hidden Markov Model (HMM), the observed data are modeled as a Markov chain plus independent noises, hence loosely speaking, the model has a short memory. In this article, we introduce a broad class of models, Quasi Hidden Markov Models (QHMMs), which incorporate long memory in the models. We develop the forward-backward algorithm and the Viterbi algorithm associated with a QHMM.

We illustrate the applications of the QHMM with the change-point problems. The structure of the QHMM enables a non-Bayesian approach. The input parameters of the model are estimated by the maximum likelihood principle. The exact inferences on change-point problems under a QHMM have a computational cost $O(T^2)$, which becomes prohibitive for large data sets. Hence we also propose approximate algorithms, which are of $O(T)$ complexity, by keeping a long but selected memory in the computation. We illustrate with step functions with Gaussian noises and Poisson processes with changing intensity. The approach bypasses model selection, and our numerical study shows that its performance is comparable and sometimes superior to the Binary Segmentation (BS) algorithm and the Pruned Exact Linear Time (PELT) method.

Autoregressive Models using Geometric Stable Distributions

Kuttykrishnan Adavalath Puthiyaveetil

Sir Syed College, India. *apkmtr@gmail.com*

Abstract: In the last two decades, there has been an increasing interest in developing the theory and applications of geometric stable distributions. The class of geometric stable distributions is a four-parameter family of distributions and this class of distributions arises as a limiting distribution of geometric random sums of independent and identically distributed random variables. Since the geometric random sums frequently appear in many applied problems in various areas, the geometric stable distributions have wide variety of applications especially in the field of reliability, biology, economics, financial mathematics etc.

In this paper we introduce and study autoregressive time series models with geometric stable distribution as the marginal distribution. The need for such models arises from the fact that many naturally occurring time series are clearly non-Gaussian and most frequently the variable of interest occur as geometric random sums of independent and identically distributed variables. As a particular case of such models we developed a first order autoregressive process with Geometric Pakes generalized asymmetric Linnik distribution as marginal distribution. Some of the properties like autocorrelation, time reversible character, sample path behavior etc. of the model are also discussed in this paper. Also we developed higher order extension of the model.

A bivariate distribution related to geometric Pakes asymmetric Laplace and Linnik distribution is introduced and bivariate time series model corresponding to this distribution is discussed. Applications of such autoregressive models in different fields are discussed in this paper.

Bayesian Local Influence Analysis for Generalized Autoregressive Conditional Heteroscedasticity Model with Empirical Applications

Hongxia Hao

Southeast University, China. *hong_xia_mm@163.com*

Processed Abstract: With the good capacity of addressing the dependency of conditional second moments of returns on financial assets, generalized autoregressive conditional heteroscedasticity (GARCH) model has been widely used in financial time series. This paper develops a Bayesian local influence analysis method for assessing the minor perturbations to the prior, individual observations, and the sampling distribution in GARCH model. A perturbation model to characterize the different perturbations is introduced and a Bayesian perturbation manifold to the perturbation model is constructed to characterize the intrinsic structure of the perturbation model by calculating geometric quantities. Several local influence measures are proposed to quantify the degree of various perturbations based on several objective functions.

Several numerical studies are conducted to evaluate the finite sample performance of the proposed method. Two empirical studies involving GARCH modeling of the continuously compounded daily returns on the New York Stock Exchange (NYSE) composite index and the series of daily log-returns for the stock index S&P500 illustrate the effectiveness of the proposed method.

On Statistical Tests for Change Points of Poisson Processes

Christian Farinetto

Université du Maine, France. *christian.farinetto@univ-lemans.fr*

Abstract: The work we present concerns is devoted to tests for Poisson processes admitting change points.

We first present the problem of detecting ruptures in the intensity function for a class of inhomogeneous Poisson point processes when a misspecified model is used in the estimation procedure. The intensity function of interest admits abrupt changes at a change point that has to be estimated from the observations. We describe the behavior of the Generalized Likelihood Ratio and Wald's tests constructed on the basis of a misspecified model in the asymptotics of large samples. We show that the Type I error rate is preserved. The power functions are studied under local alternatives and compared numerically.

In the second part of the presentation we consider the problem of testing whether realizations of a spatial Poisson process come from a circular intensity model or from an elliptical intensity model. We describe the behavior of the Generalized Likelihood Ratio and Wald's test, in the asymptotics of large samples. The power functions are studied under local alternatives and results of numerical simulations comparing them to the Neyman Pearson envelope are presented.

Tue, June 28 (08:30-10:10) | DL11 | Sponsor: Chinese Statistical Association (Taiwan)

Design of Experiments

Chair: Tsung-Chi Cheng (National Chengchi University)

Distinguished Lecturer: Ching-Shui Cheng (Academia Sinica)

Optimal Design of fMRI Experiments

Ching-Shui Cheng

Academia Sinica, Taiwan. *cheng@webmail.stat.sinica.edu.tw*

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 the comparisons of HRFs in the case of more than one type of stimuli. We establish the statistical optimality of some well-known fMRI designs, and identify several classes of new designs. Systematic methods of constructing optimal and highly efficient designs are also given. In the case of one or two types of stimuli, our results are based on a connection between fMRI designs and circulant biased weighing designs.

Experimental Designs for Functional MRI with Uncertain Model Matrix

Ming-Hung Kao

Arizona State University, United States. *mkao3@asu.edu*

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.

Optimal Design of fMRI Experiments Using Circulant (Almost-)Orthogonal Arrays

Frederick Kin Hing Phoa

Academia Sinica, Taiwan. *fredphoa@stat.sinica.edu.tw*

Abstract: Functional magnetic resonance imaging (fMRI) is a pioneering technology for studying brain activity in response to mental stimuli. Although efficient designs on these fMRI experiments are important for rendering precise statistical inference on brain functions, they are not systematically constructed. Design with circulant property is crucial for estimating a hemodynamic response function (HRF) and discussing fMRI experimental optimality. In this talk, we develop a theory that not only successfully explains the structure of a circulant design, but also provides a method of constructing efficient fMRI designs systematically. We further provide a class of two-level circulant designs with good performance (statistically optimal), and they can be used to estimate the HRF of a stimulus type and study the comparison of two HRFs. Some efficient three- and four-levels circulant designs are also provided. This is a joint work with Dr. Yuan-Lung Lin of Academia Sinica, Taiwan and Professor Ming-Hung Kao of Arizona State University, USA.

Tue, June 28 (10:30-12:10) l DL02 l Sponsor: IMS

Building Bridges: New Bayesian Insights into Old Problems

Chair: Ming-Hui Chen (University of Connecticut)

Distinguished Lecturer: Kerrie Mengersen (Queensland University of Technology)

Building Bridges: New Bayesian Insights into Old Problems Reflections on Bayesian Priors

Kerrie Mengersen

Queensland University of Technology, Australia. *k.mengersen@qut.edu.au*

Abstract: Priors are one of the tenets of Bayesian modelling, yet their formulation, properties and impact are still being understood. In this presentation we consider a range of 'real-world' problems in which a Bayesian approach has been instrumental, and the corresponding priors that have been considered. The first example involves priors for latent variable structures, namely mixtures and hidden Markov models. The second involves elicitation and formulation of priors based on virtual reality and immersive environments. The problems addressed include brain degeneration and hunting jaguars in Peru. In addition to exploring how these priors are constructed and incorporated in the models, we ask the question: how much more insight do we gain about the problem of interest through this approach to statistical analysis?

Simultaneous Estimation of Spatial Frequency Fields and Measurement Locations, with an Application to Spatial Location of Late Middle English Texts

Geoff Nicholls

University of Oxford, United Kingdom. *nicholls@stats.ox.ac.uk*

Abstract: Our data are indirect measurements of spatial frequency fields. In our application several hundred distinct multivariate frequency fields over each point in space, and we have approximately one thousand data vectors, each recording hundreds of Bernoulli observations of each frequency field at each point. If we knew the locations at which the observations had been taken we could interpolate the frequency fields (using for example, logistic regression for a spatial field). In fact the observation locations are known only for a minority of the data vectors (we call these data vectors "anchors"). We give models and joint estimation procedures for the locations of the non-anchor data vectors and the unobserved spatial frequency fields. The anchor vectors inform the spatial frequency fields in the usual way. However the non-anchors also inform fields through the model for spatial frequency fields, as the model smooths Bernoulli responses in the posterior predictive distribution for the data. The parameter space is huge and the data sparse, as we must reconstruct hundreds of spatial fields, data vector locations and observation model parameters. We apply our methods to sample-based Bayesian inference for locations of dialect samples on a map of England. The method exploits dialect-based spellings to locate these samples. The data are feature vectors extracted from written dialect samples. Just a fraction of the feature vectors ("anchors") have an associated spatial location. The data set is large, but sparse, since a given word has a large number of variant spellings which may appear in just a few documents.

Keywords: spatial statistics, complex spatial models, bayes factor, MCMC, spatial dialect fields

Joint work with Mr Ross Haines, Prof Michael Benskin

Praising the Prior

Eric-Jan Wagenmakers

University of Amsterdam, Netherlands. *EJ.Wagenmakers@gmail.com*

Abstract: The careful specification of prior distributions is important for Bayesian inference, and this is especially the case for Bayes factor hypothesis tests. I will present several examples to demonstrate that this dependence on the prior distribution can often be a strength rather than a weakness.

Tue, June 28 (10:30-12:10) | IP02 | Sponsor: IMS

Modern Developments in Multivariate Data

Organizer: Debajyoti Sinha (Florida State University)

Chair: Yiyuan She (Florida State University)

Skew-symmetric Models for Highly Skewed Clustered Data

Debajyoti Sinha

Florida State University, United States. *sinhad@stat.fsu.edu*

Abstract: Model and analysis of health-care data with highly skewed clustered response have drawn significant attention in the recent years. We propose a new class of skew-symmetric models for the highly skewed clustered response while taking into account the flexible structure for within cluster association. We introduce continuously differentiable estimating equations for the regression parameters to obtain consistent and asymptotically normal estimators with corresponding variance that can be also estimated consistently via sandwich variance estimation approach. We also present a full Bayesian procedure for our multivariate skew-symmetric models. We examine and compare the performance and robustness of our proposed methods for finite sample sizes via simulation studies. We illustrate the practical advantages of our methods via analysis of data from a real life biomedical study.

Bayes Theory and Methods for Large Networks

Debdeep Pati

Florida State University, United States. *debdeep.isi@gmail.com*

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.

Global-Local Shrinkage Priors for Variable Selection and Estimation

Malay Ghosh

University of Florida, United States. *ghoshm@stat.ufl.edu*

Abstract: The paper introduces a general class of one-group shrinkage priors for variable selection and estimation. This general class includes essentially all the priors proposed by multiple authors including the now famous "horseshoe prior". These priors are subclassified into two classes: those with exponential tails and those with polynomial tails. It is shown that priors belonging to either one of the two classes attains asymptotic consistency in the sense that the set of selected nonzero regressors coincides with the set of true nonzero regressors with probability tending to 1. However, while the subclass of polynomial tailed priors attains asymptotic efficiency in the sense that the vector of selected regressors attains asymptotic normality at the right rate, this is not so for priors with exponential tails.

Parsimonious Tensor Response Regression with Applications to Neuroimaging Analysis

Xin Zhang

Florida State University, United States. *henry@stat.fsu.edu*

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 real neuroimaging studies.

(Joint work with Dr. Lexin Li)

Tue, June 28 (10:30-12:10) | IP03 | Sponsor: IMS

Statistical Analysis of Complex Correlated Data

Organizer: Peter XK Song (University of Michigan)

Chair: Peter XK Song (University of Michigan)

Assessing Genomic Risk for Learning Problems with Neuroimaging Data

Heping Zhang

Yale University School of Public Health, United States. *heping.zhang@yale.edu*

Abstract: Modeling the link between neurocognition and brain physiology may fail to yield insightful inference without an accurate biological characterization of its relationship. Instead of directly modeling this relationship, for example, with a linear model between clinical traits and neuroimaging predictors, we consider pairwise similarity of neuroimaging features between subjects in a sample through the Similarity Model, which requires very minimal assumptions on underlying biology. From this Similarity Model, we propose a novel method to assess risk scores for neurocognitive deficits using neuroimaging data through a clustering-based algorithm. The interaction between these risk scores and genomic factors are then empirically examined to identify genomic markers having a significant impact on learning ability from the Pediatric Imaging, Neurocognition, and Genomics (PING) study. We observed there that the gene × risk score interactions improve power in a genome-wide association study. We then identify SNPs that achieving genome-wide significance for association with learning ability in samples from both the PING study and a replication study. This is a joint work with Chintan Mehta and Canhong Wen.

Simultaneous Feature Selection and Precision Matrix Estimation in High-dimensional Multivariate Regression Models

Zehua Chen

National University of Singapore, Singapore. *stachenz@nus.edu.sg*

Abstract: We consider multivariate regression models with a q-dimensional response vector, a p-dimensional feature space and a sample of size n. We deal with the problem of feature selection and precision matrix estimation of the models in the case that both q and p are large compared with the sample size n. In theoretical consideration, we allow them to diverge to infinity as n goes to infinity. We give a conditional formulation of the multivariate regression model and propose an iterated alternate method which alternates at each iteration between a feature selection step and a precision matrix estimation step. At the feature selection step, we use a sequential feature selection procedure called sequential Lasso (SLasso). At the precision matrix estimation step, we adopt the neighborhood detection approach and use a sequential scaled pairwise selection (SSPS) method. We will discuss the detailed algorithm of the iterated alternate method as well as its asymptotic properties. Simulation studies comparing the iterated alternate method with other available methods will be presented. An application to a real data set will be reported as well.

Residual-based Model Diagnosis Methods for Mixture Cure Models

Yingwei Peng

Queen's University, Canada. *pengp@queensu.ca*

Abstract: Model diagnosis, an important issue in statistical modeling, has not yet been addressed adequately for cure models. We focus on mixture cure models in this work and propose some residual-based methods to examine the fit of the mixture cure model, particularly the fit of the latency part of the mixture cure model. The new methods extend the classical residual-based methods to the mixture cure model. Numerical work shows that the proposed methods are capable of detecting lack-of-fit of a mixture cure model, particularly in the latency part, such as outliers, improper covariate functional form, or nonproportionality in hazards if the proportional hazards assumption is employed in the latency part. The methods are illustrated with two real datasets that were previously analyzed with mixture cure models.

Composite Quantile Regression for Correlated Data

Heng Lian

The University of New South Wales, Australia. *heng.lian@unsw.edu.au*

Abstract: This study investigates composite quantile regression estimation for longitudinal data on the basis of quadratic inference functions. By incorporating the correlation within subjects, our proposed CQRQIF estimator has the advantages of both robustness and high estimation efficiency for a variety of error distributions. The theoretical properties of the resulting estimators are established. Given that the objective function is non-smooth and non-convex, an estimation procedure based on induced smoothing is developed. We prove that the smoothed estimator is asymptotically equivalent to the original estimator. We also propose the weighted composite quantile regression estimation to improve the estimation efficiency further in some situations. Extensive simulations are conducted to compare different estimators, and a real data analysis is used to illustrate their performances.

Tue, June 28 (10:30-12:10) I IP44

Stein's Method and Random Graphs

Organizer: Adrian Roellin (National University of Singapore)

Chair: Adrian Roellin (National University of Singapore)

Change Point Detection in Evolving Network Models

Shankar Bhamidi

The University of North Carolina at Chapel Hill, United States. *bhamidi@email.unc.edu*

Abstract: The last few years have seen an explosion in the amount of data on real world networks, including networks that evolve over time. A number of mathematical models have been proposed to understand the evolution of such networks and explain the emergence of a wide array of structural features such as heavy tailed degree distribution and small world connectivity of real networks. In this paper we consider one famous class of such models, the preferential attachment model. We formulate and study the regime where the network transitions from one evolutionary scheme to another. In the large network limit we derive asymptotics for various functionals of the network including degree distribution and maximal degree. We study functional central limit theorems for the evolution of the degree distribution which feed into proving consistency of a proposed estimator of the change point.

Rates of Convergence for Multivariate Normal Approximation with Applications to Dense Graphs

Xiao Fang

National University of Singapore, Singapore. *stafx@nus.edu.sg*

Abstract: We provide a new general theorem for multivariate normal approximation on convex sets. The theorem is formulated in terms of a multivariate extension of Stein couplings. We apply the results to a homogeneity test in dense random graphs. This talk is based on joint work with Adrian Roellin.

Limit Behavior of Some Polya Urn Models Associated to Preferential Attachment Graphs and Random Trees

Nathan Ross

The University of Melbourne, Australia. *nathan.ross@unimelb.edu.au*

Abstract: I will discuss a family of Polya urn models that arise when studying degree statistics in preferential attachment graphs and lengths of spanning subtrees in some generative tree models such as Remy's algorithm for binary trees. The limiting distributions of the number of balls of a given color in the urn varies considerably across the family and can be difficult to describe. However, these limits are fixed points of certain probabilistic distributional transformations and this perspective provides proofs of convergence, sometimes with rates via Stein's method, and leads to further properties of the limits. Joint work with Erol Pekoz and Adrian Roellin.

Bounds on the Condensation Threshold in Stochastic Block Models

Joe Neeman

University of Texas at Austin / University of Bonn, United States / Germany. *neeman@iam.uni-bonn.de*

Abstract: Consider a random graph model consisting of k communities of equal size. Edges are added to the graph independently at random, but are somewhat more likely to connect two vertices that belong to the same community. We choose parameters so that the average degree of the graph is fixed as the number of vertices tends to infinity.

This random graph model is believed to have two phase transitions as the strength of community attachment increases. At the first transition, it becomes information-theoretically possible to detect the hidden communities from the graph structure; at the second, one can find them in polynomial time. We give upper and lower bounds on this first phase transition, which is known as the "condensation threshold" in statistical physics. Our bounds are asymptotically sharp in some limits, but not in others.

This is joint work with Jess Banks, Cristopher Moore, and Praneeth Netrapalli.

Tue, June 28 (10:30-12:10) l IP18 l Sponsor: Korea

New Frontiers of Longitudinal/Clustered Data Analysis

Organizer: Mi-Ok Kim (Cincinnati Children's Hospital Medical Center)

Chair: Yehua Li (Iowa State University)

Disease Progress Monitoring Based on Bayesian Joint Modeling of Left-truncated Longitudinal and Survival Outcomes

Mi-Ok Kim

Cincinnati Children's Hospital Medical Center, United States. *MiOk.Kim@cchmc.org*

Abstract: This talk proposes a Bayesian joint modeling approach that accounts for left-truncation and asymmetric serially correlated longitudinal measurements. Joint modeling approaches have been introduced to model informative missing observations and drop outs in the longitudinal outcome analysis and account for time-varying covariates in the survival outcome analysis. A real data example from Cystic Fibrosis (CF) Foundation patient registry data from year 1996 to 2009 shows that longitudinal pulmonary function outcomes were measured among the survivors over different age periods and are serially correlated with the marginal distribution highly skewed. These features, commonly found in registry based data, have not been adequately accommodated in the joint modeling analysis. The proposed method uses dynamic modeling to account for the serial correlation and facilitate the forecasting.

A Shared Parameter Model for Curve Registration in the Presence of Informative Dropout

Sarah J. Ratcliffe

University of Pennsylvania, United States. *sratclif@upenn.edu*

Abstract: Longitudinal outcomes are often measured repeatedly over time until an event of interest occurs. When specific types of longitudinal trajectories are more likely to drop out of the study, it may be desirable to consider a model for the joint distribution of the longitudinal and event time outcomes. A common longitudinal outcome in Women's Health studies is the labor curve, cervical dilation measured over time during stage 1 of labor. In women attempting Vaginal Birth After Cesarean (VBAC), uterine rupture may occur resulting in early termination of observation of the labor curve due to cesarean delivery. It is likely that specific features of the start and pace of labor, which can be captured by random effects via curve registration, are predictive of dropout in the sense that curves with slower paces and earlier start times are more likely to be terminated early. Here, a shared parameter model is developed where curve registration random effects are used to link the longitudinal and dropout models.

Marginal Zero-inflated Regression Models for Cross-sectional and Clustered Count Data

Daniel Hall

University of Georgia, United States. *danhall@uga.edu*

Abstract: Count data with more zeros than predicted by standard count distributions are commonly analyzed via zero-inflated (ZI) regression models such as the ZI Poisson (ZIP), ZI binomial (ZIB), and ZI negative binomial (ZINB) models. These models assume the data are generated from a mixture of a degenerate distribution at zero and a standard count distribution such as the Poisson, binomial or negative binomial distribution. Covariate effects are modeled via generalized linear model type specifications for the mixing probability p and the non-degenerate component's mean, λ. Although the mixture formulation is appealing for many problems, if interest centers on the marginal mean response $\mu = (1-p)\ \lambda$, interpretation of regression parameters is awkward because covariate effects operate on the marginal mean indirectly through p and λ. To alleviate this problem we propose marginal ZI models where covariate effects on μ are modeled directly via a log linear regression equation. Computational methods for fitting these models via the EM algorithm and obtaining suitable standard errors are proposed. The models are illustrated on real data and small sample properties of Wald inferences for these models are evaluated via simulation. The models are extended to the clustered (e.g., longitudinal) data context via mixed-effect formulations and we show how such mixed-effect versions can be marginalized over the random effects distribution to yield parameters with both subject-specific and population averaged interpretations. This marginalization is much simpler and more appealing than for traditional ZI and hurdle regression models with mixed-effects.

Doubly Robust Generalized Estimating Equations with Consistent Variance Estimator when One Auxiliary Model is Misspecified

Soeun Kim

The University of Texas Health Science Center at Houston, United States. *Soeun.Kim@uth.tmc.edu*

Abstract: The doubly-robust generalized estimating equation (DR-GEE) is a popular analytic tool for repeated measurements with missing data. It requires two assumptions on auxiliary models for outcome and the observation indicator, and produces a consistent point estimator when either of the model assumptions is correct. Another advantage is an easy variance formula due to asymptotic independence between the estimator of interest and the estimators from the auxiliary models. However, this feature can be capitalized only when both models are correctly specified. In this paper, we propose an alternative DR-GEE that produces a consistent variance estimator even when one of the auxiliary models is misspecified. The main idea is to construct estimating equations for the auxiliary models so that the estimators of main interest and the estimators from auxiliary models are asymptotically independent. We illustrate the method on data from a clinical trial for hypertension.

Tue, June 28 (10:30-12:10) l IP38 l Sponsor: Australia and New Zealand

Dimension Reduction and Variable Selection for High-Dimensional Data

Organizer: Inge Koch (The University of Adelaide)

Chair: Valentin Patilea (ENSAI)

Pseudo Sufficient Dimension Reduction and Variable Selection

Xiangrong Yin

University of Kentucky, United States. *yinxiangrong@gmail.com*

Abstract: Sufficient dimension reduction has achieved great success in recent years. When the sample covariance matrix of the predictors is not invertible, many sufficient dimension reduction methods take an ad hoc ridge regression approach. A question that has been raised for a long while is whether such an estimator is still in the central subspace. In this paper, we propose new concepts of pseudo sufficient dimension reduction and pseudo sufficient variable selection to answer this question. Based on an underlying relationship between the ridge regression and the measurement error regression, we propose a general procedure to obtain pseudo estimates. Using an ensemble idea, our proposed pseudo estimates are better than the traditional estimate or a ridge estimate for highly correlated predictors. Large p small n issue is discussed, while theoretical properties are obtained. Simulation studies and two real data analyses are used to demonstrate the advantage of our methods.

On the Effective Number of Principal Components in HDLSS Context

Sungkyu Jung

University of Pittsburgh, United States. *sungkyu@pitt.edu*

Abstract: This paper deals with the problem of how many principal components to retain in the high dimension, low sample size context. Assuming that the first m eigenvalues of the covariance matrix are diverging, we show that even though estimated principal component directions are inconsistent, the corresponding scores contain a good amount of information on the population scores. To estimate the number of components m, we propose to test sequential null hypotheses Hk : m = k for increasing values of k. Our estimate of m is the smallest k for which Hk is accepted. The test statistics are based on the lengths of residual scores that are obtained by removing the first k principal components. We show that under the global null hypothesis, H0 : m = 0, the squared lengths are asymptotically normal, and under Hk, k ≥ 1, they are asymptotically left-skewed as the dimension grows. The proposed test procedure performs well in high-dimensional simulation studies, and provides reasonable estimates of m in a number of real data examples.

Kernel Naive Bayes for High Dimensional Pattern Recognition

Kanta Naito

Shimane University, Japan. *naito@riko.shimane-u.ac.jp*

Abstract: Kernel Naive Bayes (KNB) for pattern recognition in the setting of High Dimensional Low Sample Size (HDLSS) is proposed, which is an effective combination of the usual Naive Bayes and the kernel method. As a variant of KNB, Smoothed KNB (SKNB) is also considered, which is based on expectations of matrices that appear in KNB. Asymptotic theory as the dimension grows is developed, and numerical experiments and applications to real HDLSS data sets reveal that KNB and SKNB are shown to work well.

Classification and Variable Selection for High-dimensional Data with Application to Proteomics

Inge Koch

The University of Adelaide, Australia. *inge.koch@adelaide.edu.au*

Abstract: For two-class classification problems Fisher's discriminant rule performs well provided the dimension is smaller than the sample size. As the dimension increases, Fisher's rule may no longer be adequate, and can perform as poorly as random guessing. For high-dimension low sample size (HDLSS) data, dimension reduction and feature selection have become essential prior to applying any classification rule.

In this talk we look at different ways of incorporating feature selection into Fisher's classification rule and the naïve Bayes rule including the 'Features Annealed Independence Rule' (FAIR) of Fan and Fan (2008), and the 'Naïve Canonical Correlation' approach of Tamatani, Koch and Naito (2012). We examine the behavior of such rules and look at asymptotic properties as the dimension and sample size grow.

Proteomics is a rapidly growing research area within bioinformatics which focuses on identification of proteins, peptides and biomarkers from peptide concentrations. We consider proteomics imaging mass spectrometry data – HDLSS data with an underlying spatial distribution – here from patients with endometrial cancer. For these data we examine the performance of feature selection and classification rules in predicting which patients' cancer will metastasise.

Tue, June 28 (10:30-12:10) | IP54

Recent Advances in Analysis of Complex Data

Organizer: Xin Yuan Song (The Chinese University of Hong Kong)

Chair: Xin Yuan Song (The Chinese University of Hong Kong)

Bayesian Neural Networks for Personalized Medicine

Faming Liang

University of Florida, United States. *faliang@ufl.edu*

Abstract: Complex diseases such as cancer have often heterogeneous responses to treatment, and this has attracted much interest in developing individualized treatment rules to tailor therapies to an individual patient according to the patient-specific characteristics. In this talk, we discuss how to use Bayesian neural networks to achieve this goal, including how to select disease related features. The theoretical properties of Bayesian neural networks is studied under the small-n-large-P framework, and simulation is done using the parallel stochastic approximation Monte Carlo algorithm on a multi-core computer. The performance of the proposed approach is illustrated via simulation studies and a real data example.

A Concave Pairwise Fusion Approach to Subgroup Analysis

Jian Huang

The University of Iowa, United States. *jian-huang@uiowa.edu*

Abstract: An important step in developing individualized treatment strategies is to correctly identify subgroups of a heterogeneous population, so that specific treatment can be given to each subgroup. In this paper, we consider the situation with samples drawn from a population consisting of subgroups with different means, along with certain covariates. We propose a penalized approach for subgroup analysis based on a regression model, in which heterogeneity is driven by unobserved latent factors and thus can be represented by using subject-specific intercepts. We apply concave penalty functions to pairwise differences of the intercepts. This procedure automatically divides the observations into subgroups. We develop an alternating direction method of multipliers algorithm with concave penalties to implement the proposed approach. We also establish the theoretical properties of our proposed estimator and determine the order requirement of the minimal difference of signals between groups in order to recover them. These results provide a sound basis for making statistical inference in subgroup analysis. Our proposed method is further illustrated by simulation studies and analysis of the Cleveland heart disease dataset. This is joint work with Shujie Ma.

Bayesian Shape-restricted Analysis of Complex Data using Gaussian Process Priors

Taeryon Choi

Korea University, South Korea. *trchoi@korea.ac.kr*

Abstract: In this paper, we propose a Bayesian method to estimate shape-restricted functions using Gaussian process priors for complex data. The proposed model enforces shape-restrictions by assuming that the derivatives of the functions are squares of Gaussian processes. The resulting functions after integration are monotonic, monotonic convex or concave, U--Shaped, and S--shaped. We illustrate the empirical performance of the proposed models with synthetic and real data, and discuss their potentials and extensions.

Ball Divergence: Nonparametric Two Sample Test

Xueqin Wang

Sun Yat-sen University, China. *wangxq88@mail.sysu.edu.cn*

Abstract: In this paper, we first introduce Ball Divergence, a novel measure of the difference between two probability measures in finite dimensional Banach space, and show that the Ball Divergence of two probability measures is zero if and only if these two probability measures are identical. Using Ball Divergence, we present a metric rank test procedure to detect the equality of distribution measures underlying independent samples. We show that this multivariate two sample test statistic is consistent with the Ball Divergence, and it converges to a mixture of Chi-square distributions under the null hypothesis and a normal distribution under the alternative hypothesis. Importantly, we prove its consistency against a general alternative hypothesis. Even without the moment assumption, the test based on Ball Divergence is robust to the heavy-tail data. Moreover, the result does not depend on the ratio of the two imbalanced sample sizes, ensuring that the test is robust and can be applied to imbalanced data. Numerical studies confirm that our test is superior to several existing tests in terms of Type I error and power. We conclude our paper with two applications of our method: one is for virtual screening in drug development process and the other is for genome wide expression analysis in hormone replacement therapy.

Tue, June 28 (10:30-12:10) l TCP18

Recent Advances in Statistical Genetics

Organizer: Hong Zhang (Fudan University)

Chair: Zhaohai Li (The George Washington University)

A Principal Score Test for Analyzing Multiple Genetic Markers

Jinbo Chen

University of Pennsylvania, United States. *jinboche@mail.med.upenn.edu*

Abstract: Testing multiple genetic markers simultaneously for association has been well appreciated as an approach complimentary to tests of individual markers. Here, we propose a new test statistic based on aggregation of likelihood score functions for testing marginal associations of each individual marker. Our method has close connections to the principal component regression method, and has competitive power relative to some widely-used gene-based association tests. Through extensive computer simulations and application to real genetic association studies, we show that our method is powerful when the genetic markers are in moderate to strong linkage disequilibrium.

Joint work with Zhengbang Li, Kai Yu, Tianxi Cai

Extension of the Peters-Belson Method to Estimate Health Disparities Among Multiple Groups using Logistic Regression with Survey Data

Yan Li

University of Maryland, United States. *yli6@umd.edu*

Abstract: Determining the extent of a disparity, if any, between groups of people, is of interest in many fields, including public health for medical treatment and prevention of disease. An observed difference in the mean outcome between an advantaged group (AG) and disadvantaged group (DG) can be due to differences in the distribution of relevant covariates. The Peters-Belson (PB) method fits a regression model with covariates to the AG t0 predict, for each DG member, their outcome measure as if they had been from the AG. The difference between the mean predicted and observed outcomes of DG members is the (unexplained) disparity of interest. We focus on applying the PB method to estimate the disparity based on logistic regression models using data collected from complex surveys with multiple DGs. Estimators of the unexplained disparity, an analytic variance estimator that is based on the Taylor linearization variance method, as well as a Wald test for testing a joint null hypothesis of zero for unexplained disparities between multiple minority groups and a majority group, are provided. Simulation studies and analyses of disparity in BMI in NHANES 1999-2004 are conducted.

A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations

Jianxin Shi

National Cancer Institute, United States. *jianxin.shi@nih.gov*

Abstract: The central challenge in tumor sequencing studies is to identify driver genes and pathways, investigate their functional relationships and nominate drug targets. The efficiency of these analyses, particularly for infrequently mutated genes, is compromised when subjects carry different combinations of driver mutations. Mutual exclusivity analysis helps address these challenges. To identify mutually exclusive gene sets (MEGS), we developed a powerful and flexible analytic framework based on a likelihood ratio test and a model selection procedure. Extensive simulations demonstrated that our method outperformed existing methods for both statistical power and the capability of identifying the exact MEGS, particularly for highly imbalanced MEGS. Our method can be used for de novo discovery, pathway-guided searches or for expanding established small MEGS. We applied our method to the whole exome sequencing data for thirteen cancer types from The Cancer Genome Atlas (TCGA). We identified multiple previously unreported non-pairwise MEGS in multiple cancer types. For acute myeloid leukemia, we identified a MEGS with five genes (FLT3, IDH2, NRAS, KIT and TP53) and a MEGS (NPM1, TP53and RUX1) whose mutation status was strongly associated with survival (P=6.7×10-4). For breast cancer, we identified a significant MEGS consisting of TP53 and four infrequently mutated genes (ARID1A, AKT1, MED23 and TBL1XR1), providing support for their role as cancer drivers.

Strategies for Conducting Multi-locus Analysis using Single-locus Summary Statistics

Kai Yu

National Cancer Institute, United States. *yuka@mail.nih.gov*

Abstract: To maximize the chance of finding outcome-associated variants, a number of consortia have been formed to conduct meta-analysis integrating results from multiple genome-wide association studies (GWAS). The GWAS meta-analysis typically combines summary statistics from participating studies on each single nucleotide polymorphisms (SNP) by using a fixed effect model. As SNP-level summary statistics (e.g., odds ratio estimates, p-values) generated by consortia become increasingly available, there is a great interest to conduct mulita-locus analysis based on the easily accessible SNP-level summary statistics, instead of the individual-level GWAS data, in order to search the genome more thoroughly for outcome-associated loci with relatively small marginal effect. In this talk, we will describe several strategies for dealing with challenges arising from applications of this type of analysis on summary statistics generated from large consortia.

Tue, June 28 (10:30-12:10) | TCP19

Recent Developments in Model Checking

Organizer: Xu Guo (Nanjing University of Aeronautics and Astronautics)

Chair: Maolin Pan (Nanjing University)

Enhancements of Nonparametric Generalized Likelihood Ratio Test: Bias-correction and Dimension Reduction

Xu Guo

Nanjing University of Aeronautics and Astronautics, China. *liushengjunyi@163.com*

Abstract: Nonparametric generalized likelihood ratio test is popularly used for model checking for regressions. However, there are two issues that may be the barriers for its powerfulness. First, the bias term in its liming null distribution causes the test not to well control type I error and thus Monte Carlo approximation for critical value determination is required. Second, it severely suffers from the curse of dimensionality due to the use of multivariate nonparametric function estimation. The purpose of this paper is thus two-fold: a bias-correction is suggested to this test and a dimension reduction-based model-adaptive enhancement is recommended to promote the power performance. The proposed test still possesses the Wilks phenomenon, and the test statistic can converge to its limit at a much faster rate and is much more sensitive to alternative models than the original nonparametric generalized likelihood ratio test as if the dimension of covariates were one. Simulation studies are conducted to evaluate the finite sample performance and to compare with other popularly used tests. A real data analysis is conducted for illustration.

A Robust Adaptive-to-model Enhancement Test for Parametric Single-index Models

Cui Zhen Niu

Renmin University of China, China. *nczlbc_890@126.com*

Abstract: This paper is devoted to check when there are outliers in observations, whether the underlying model is of parametric single-index structure. The purpose of this paper is two-fold. First, a test that is robust against outliers is suggested. The Hampel's second-order influence function of the test statistic is proved to be bounded. Second, the test fully uses the dimension reduction structure of the hypothetical model and automatically adapts to alternative models when the null hypothesis is false. Thus, the test can greatly overcome the dimensionality problem and is still omnibus against general alternative models. The performance of the test is demonstrated by both Monte Carlo simulation studies and an application to a real dataset.

Dimension Reduction-based Model Checking for Survival Models

Jingke Zhou

Ningbo University, China. *zhoujingke1118@163.com*

Abstract: Model specification test is investigated for survival model with high dimensional covariates, while most existing methods suffer from typical curse of dimensionality since they all involve multivariate nonparametric regression estimation. To alleviate the effect of curse of dimensionality, dimension reduction technique is introduced in this paper. Theoretically, the consistency under null hypothesis and alternative hypothesis are proved. For finite sample setting, bootstrap strategy is employed to approximate the critical values of the test. Finally, numerical simulations and a real data analysis are conducted to show the performance of the proposed test.

An Adaptive-to-model Test for Partially Parametric Single-index Models

Xuehu Zhu

Xi'an Jiaotong University, China. *zhuxuehu0320@163.com*

Abstract: Residual marked empirical process-based tests are commonly used in regression models. However, they suffer from data sparseness in high-dimensional space when there are many covariates. This paper has three purposes. First, we suggest a partial dimension reduction adaptive-to-model testing procedure that can be omnibus against general global alternative models although it fully uses the dimension reduction structure under the null hypothesis. This feature is because that the procedure can automatically adapt to the null and alternative models, and thus greatly overcomes the dimensionality problem. Second, to achieve the above goal, we propose a ridge-type eigenvalue ratio estimate to automatically determine the number of linear combinations of the covariates under the null and alternatives. Third, a Monte-Carlo approximation to the sampling null distribution is suggested. Unlike existing bootstrap approximation methods, this gives an approximation as close to the sampling null distribution as possible by fully utilising the dimension reduction model structure under the null. Simulation studies and real data analysis are then conducted to illustrate the performance of the new test and compare it with existing tests.

Tue, June 28 (10:30-12:10) | TCP10

Bayesian Modeling with Complex Survey Data

Organizer: Cici Chen Bauer (Brown University)

Chair: Tony Sit (The Chinese University of Hong Kong)

Robust Bayesian Models for Surveys with Missing Data and External Information

Sahar Zangeneh

Fred Hutchinson Cancer Research Center, United States. *saharzz@fhcrc.org*

Abstract: Survey data are often collected according to some complex probability sampling design. Design-based inference aims at obtaining unbiased, or minimally biased, weighted estimates of finite population quantities such as means or totals, with respect to the sampling design. Within this framework, calibration methods are often employed to incorporate external information. These methods re-adjust the original weights to satisfy constraints imposed by the external information while deviating minimally from the original weights. However, such estimators are sensitive to the choice of distance measure. An alternative paradigm is Bayesian model-based estimation, where a model is chosen to predict the non-observed data. Such models need to incorporate features of the sampling design, like sampling weights and clustering, to be robust to model misspecification. We describe two applications of this approach to survey problems involving external information: (i) probability proportional to size sampling, where aggregate information is available for sizes of non-sampled units, and (ii) survey nonresponse, when there is external information for post-stratification. Simulation comparisons with standard "design-based" methods suggest superior frequentist properties for the Bayesian method.

Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question

Guo-Liang Tian

The University of Hong Kong, Hong Kong. *gltian@hku.hk*

Abstract: Although the item count technique (ICT) is useful in surveys with sensitive questions, privacy of those respondents who possess the sensitive characteristic of interest may not be well protected due to a defect in its original design. In this talk, we propose two new survey designs (namely the Poisson ICT and negative binomial ICT) which replace several independent Bernoulli random variables required by the original ICT with a single Poisson or negative binomial random variable, respectively. The proposed models not only provide closed form variance estimate and confidence interval within [0, 1] for the sensitive proportion, but also simplify the survey design of the original ICT. Most importantly, the new designs do not leak respondents' privacy. Empirical results show that the proposed techniques perform satisfactorily in the sense that it yields accurate parameter estimate and confidence interval. (This is a joint work with Man-Lai TANG, Qin WU, and Yin LIU).

Bayesian Empirical Likelihood Methods for Complex Surveys

Changbao Wu

University of Waterloo, Canada. *cbwu@uwaterloo.ca*

Abstract: Bayesian pseudo empirical likelihood intervals for complex surveys are studied by Rao and Wu (JRSSB, 2010). Their proposed approach focuses on the finite population mean and is not applicable to multiple parameters defined through estimating equations. In this talk we present a Bayesian empirical likelihood method for complex surveys. Our proposed approach provides Bayesian inferences on multiple parameters with valid frequentist properties under the traditional design-based framework. The method covers single stage unequal probability sampling designs with negligible sampling fractions and multi-stage sampling designs with negligible sampling fractions for selecting first stage clusters. We demonstrate the usefulness of the method through regressions analysis using survey data.

Bayesian Spatial Hierarchical Models for Small Estimation with Complex Survey Designs

Cici Chen Bauer

Brown University, United States. *cici_bauer@brown.edu*

Abstract: Spatial hierarchical models have shown to be beneficial for small area estimation. However, the sampling weights that are required to reflect complex surveys are rarely considered in these models. In this talk, I will describe a method for incorporating the sampling weights for binary data when estimating, for example, small area proportions or predicting small area counts. Spatial hierarchical random effects are shown to be beneficial, with computation carried out using the integrated nested Laplace approximation, which is fast. Simulation results will be presented to show that estimation of mean squared error can be reduced when compared with more standard approaches. Bias reduction occurs through the incorporation of sampling weights, with variance reduction being achieved through hierarchical smoothing. The application of our proposed method with data taken from the Washington 2006 Behavioral Risk Factor Surveillance System will also be presented.

Tue, June 28 (10:30-12:10) l CP03

Machine Learning Session 3

Chair: Yoonkyung Lee (The Ohio State University)

On a General Procedure for Constructing Confidence Sets under Partially Identified Models

Han Jiang

The University of Hong Kong, Hong Kong. *shirleyjiang@hku.hk*

Abstract: Recent years have seen a growing body of literature focusing on partially identified models, where observable data and credible assumptions can only identify the parameter of interest with a set, called the identified set, rather than a singleton. Manski (2003) provides a recipe of partial identification problems which may not be readily amenable to conventional statistical approaches. For certain partially identified models such as those defined using moment inequalities, new algorithms have been developed for constructing confidence sets for identified sets. The problem remains, however, unsolved outside such restrictive contexts. In the present study we propose a general confidence procedure which not only generalizes existing algorithms but also finds applications to settings as yet unexplored. The main thrust of our proposed procedure lies in an expansion step, by which we construct a family of nested sets and perform inferences centred on an expansion index. Wide choices of expansion indices and their corresponding nested sets allow of the flexibility and applicability necessary for dealing with a more general class of partially identified models. To illustrate the generality of our procedure, a simulation study is presented concerning the least quantile of squares in a partially identified model setting, a problem to which no solution has yet been found in the literature.

Robust Non-convex Penalized Linear Regression with Algorithmic and Statistical Convergence

Shota Katayama

Tokyo Institute of Technology, Japan. *katayama.s.ad@m.titech.ac.jp*

Abstract: Non-convex penalized linear regression such as SCAD and MCP is an useful tool in recent statistical applications. However, standard methods using the squared L^2 loss are not robust against outliers. To handle outliers, we add parameters whose non-zero elements correspond to outliers to the standard linear regression model, and then estimate coefficients and outlier parameters with non-convex sparse penalties. Since the optimization problem considered here is totally non-convex, it suffers from many local optima and it is difficult to find a good local solution. In this talk, we shall provide an algorithmic and statistical convergence to the true coefficients on a solution that a practical algorithm outputs.

Adaptive Kernel-based FPCA for Functional Generalized Linear Models

Guangbao Guo

Shandong University, China. *ggb11111111@163.com*

Abstract: In the study, we consider adaptive kernel to functional principal component analysis (FPCA) in functional generalized linear models. We first give the properties of adaptive kernel-based FPCA after we provide the procedure of proposed approach. Some simulations are given to in order to confirm the effect of the approach.

Tue, June 28 (13:30-15:10) l DL12 l Sponsor: Hong Kong

Nonstationary Time Series: Past, Present and Beyond

Chair: Wai Keung Li (The University of Hong Kong)

Distinguished Lecturer: Ngai Hang Chan (The Chinese University of Hong Kong)

Nonstationary Time Series: Past, Present, and Beyond

Ngai Hang Chan

The Chinese University of Hong Kong, Hong Kong. *nhchan@sta.cuhk.edu.hk*

Abstract: This talk reviews recent developments of nonstationary and long-memory time series. The underlying theme of recent endeavour arises from the consideration of the order of magnitude of the observed Fisher's information number. By means of a simple AR(1) model, the so-called "SNoTE", it is shown how this number affects the nonstationary behavior in a subtle and important way. The talk elaborates some of the past and present important developments of nonstationary time series from a historical perspective. This talk concludes with discussions of some of the fascinating results involving negative moment bounds of the observed Fisher's information number, which bear important applications to multi-step ahead predictions.

Multiple-output Quantile Regression: a Survey

Marc Hallin

Université Libre de Bruxelles, Belgium. *mhallin@ulb.ac.be*

Abstract: Quantile regression is about estimating the quantiles of some *d*-dimensional response **Y** conditional on the values $\mathbf{x} \in R^p$ of some covariates **X**. The problem is well understood when d=1 (single-output). However, a response seldom comes as an isolated quantity, and **Y**, in most situations of practical interest, takes values in R^d, with $d \geq 2$ (multiple-output case). An extension to $d \geq 2$ of quantile regression concepts and methods is thus extremely desirable. Unfortunately, they all exploit the canonical ordering of the real line *R*. Such an ordering no longer exists when $d \geq 2$. As a consequence, (location and regression) quantiles, but also equally basic univariate concepts such as distribution functions, signs, ranks—all playing a fundamental role in statistical inference—do not straightforwardly extend to higher dimensions. In this talk, we discuss the various proposals that have been made in the literature, along with some new ones, and their relation to statistical depth.

Bootstrap Unit Root Inference for Non-stationary Linear Processes driven by Infinite Variance Innovations

Giuseppe Cavaliere

University of Bologna, Italy. *giuseppe.cavaliere@unibo.it*

Abstract: The contribution of this paper is two-fold. First, we derive the asymptotic null distribution of the familiar augmented Dickey-Fuller [ADF] statistics in the case where the shocks follow a linear process driven by infinite variance innovations. We show that these distributions are free of serial correlation nuisance parameters but depend on the tail index of the infinite variance process. These distributions are shown to coincide with the corresponding results for the case where the shocks follow a finite autoregression, provided the lag length in the ADF regression satisfies the same o($T^{1/3}$) rate condition as is required in the finite variance case. In addition, we establish the rates of consistency and (where they exist) the asymptotic distributions of the ordinary least squares sieve estimates from the ADF regression. Given the dependence of their null distributions on the unknown tail index, our second contribution is to explore sieve wild bootstrap implementations of the ADF tests. Under the assumption of symmetry, we demonstrate the asymptotic validity (bootstrap consistency) of the wild bootstrap ADF tests. This is done by establishing that (conditional on the data) the wild bootstrap ADF statistics attain the same limiting distribution as that of the original ADF statistics taken conditional on the magnitude of the innovations.

Tue, June 28 (13:30-15:10) l IP05 l Sponsor: IMS

Recent Advances in Large-Scale Inference

Organizer: Zheng Tracy Ke (The University of Chicago)

Chair: Yingying Wei (The Chinese University of Hong Kong)

High Dimensional Minimum Variance Portfolio Estimation

Yingying Li

The Hong Kong University of Science and Technology, Hong Kong. *yyli@ust.hk*

Abstract: We study the estimation of high dimensional minimum variance portfolio (MVP). Two settings are considered: the low frequency setting where returns are modeled as i.i.d., and the high frequency setting where returns can exhibit heteroskedasticity and possibly be contaminated by microstructure noise. We first propose an estimator of the minimum variance, which provides a benchmark for comparison. More importantly, under some sparsity assumptions on the precision matrix, we propose an estimator of MVP, which asymptotically achieves the minimum variance. Simulation and empirical studies demonstrate that our proposed portfolio performs favorably.

Detecting Rare and Weak Spikes in Large Covariance Matrices

Zheng Tracy Ke

The University of Chicago, United States. *zke@galton.uchicago.edu*

Abstract: Given n iid samples of a p-dimensional Gaussian random vector, we are interested in testing a null hypothesis where the covariance matrix Sigma is the identity against an alternative hypothesis where Sigma is a spike matrix; i.e., all eigenvalues are 1, except for r of them are larger than 1 (i.e., spiked eigenvalues). We propose CuSum and Higher Criticism as two new tests, and also investigate a trace-based test and the Tracy-Widom (TW) test (Johnstone, 2001).

We consider a Rare/Weak setting where the spikes are both sparse and individually weak (i.e., 1<< r<< p and each spiked eigenvalue is only larger than 1 by a small amount). We discover the following phase transition: the two-dimensional phase space calibrating the spike sparsity and strengths partitions into the "Region of Impossibility" and the "Region of Possibility". In Region of Impossibility, all tests are (asymptotically) powerless in separating the alternative and the null. In Region of Possibility, both the CuSum test and the trace-based test have asymptotically full power. The TW test, however, may be powerless in a sub-region of the Region of Possibility.

As a new use of the proposed tests, we derive new bounds for all individual eigenvalues as well as for all the cumulative sums of eigenvalues, both under the alternative. The former is an extension of Baik, Ben Arous and Peche (2005), where (a) the bounds are for all eigenvalues but not only for edge eigenvalues, and (b) the number of spikes is now allowed to grow with p.

The study requires careful analysis of the L1-distance and sophisticated Radom Matrix Theory. Our technical devises include (a) a Gaussian proxy model, (b) Le Cam's comparison of experiments, and (c) large deviation bounds on the empirical eigenvalues.

Neyman-Pearson Classification under High-dimensional Settings

Yang Feng

Columbia University, United States. *yangfeng@stat.columbia.edu*

Abstract: Most existing binary classification methods target on the optimization of the overall classification risk and may fail to serve some real-world applications such as cancer diagnosis, where users are more concerned with the risk of misclassifying one specific class than the other. Neyman-Pearson (NP) paradigm was introduced in this context as a novel statistical framework for handling asymmetric type I/II error priorities. It seeks classifiers with a minimal type II error and a constrained type I error under a user specified level. This article is the first attempt to construct classifiers with guaranteed theoretical performance under the NP paradigm in high-dimensional settings. Based on the fundamental Neyman-Pearson Lemma, we used a plug-in approach to construct NP-type classifiers for Naive Bayes models.

The proposed classifiers satisfy the NP oracle inequalities, which are natural NP paradigm counterparts of the oracle inequalities in classical binary classification. Besides their desirable theoretical properties, we also demonstrated their numerical advantages in prioritized error control via both simulation and real data studies.

A CLT for Random Sesquilinear Forms with Applications in RMT

Zhonggen Su

Zhejiang University, China. *suzhonggen@zju.edu.cn*

Abstract: In this talk we report a joint central limit theorem for random sesquilinear forms, which is a extension of central limit theory on random quadratic forms. As applications, we study extremal eigenvalues of large dimensional random matrices with spiked population. In particular, we find the joint distribution of grouped extremal sample eigenvalues under certain conditions. This talk is based on a joint work with Qiwen Wang and Jianfeng Yao (EJP, 2014).

Tue, June 28 (13:30-15:10) l IP19 l Sponsor: Korea

Recent Advances in Bayesian Functional Data Analysis

Organizer: Taeryon Choi (Korea University)

Chair: Taeryon Choi (Korea University)

Functional Regression Approximate Bayesian Computation for Gaussian Process Density Estimation

David John Nott

National University of Singapore, Singapore. *standj@nus.edu.sg*

Abstract: A novel Bayesian nonparametric method for hierarchical modelling on a set of related density functions is considered, where grouped data in the form of samples from each density function are available. Borrowing strength across the groups is a major challenge in this context. To address this problem, we introduce a hierarchically structured prior, defined over a set of univariate density functions, using convenient transformations of Gaussian processes. Inference is performed through approximate Bayesian computation (ABC), via a novel functional regression adjustment approach. The performance of the proposed method is illustrated via a simulation study and an analysis of rural high school exam performance in Brazil. This is joint work with Guilherme Rodrigues and Scott Sisson.

Clustering Functional Data using Principal Curve Methods

Bo Wang

University of Leicester, United Kingdom. *bo.wang@leicester.ac.uk*

Abstract: Clustering for functional data has received increasing interests in various disciplines, and a number of different methods for this purpose have been developed. We introduce a clustering method for functional data based on the principal curve clustering approach. The functional data are first decomposed using functional principal component analysis (FPCA) for dimension reduction and a model based on principal curves is then developed for clustering the corresponding principal scores. The usefulness of the proposed method is demonstrated by simulated examples and real data.

Bayesian Spectral Analysis Models for Functional Clustering

Minjung Kyung

Duksung Women's University, South Korea. *mkyung@duksung.ac.kr*

Abstract: This paper presents a Bayesian analysis for clustering of functional data. We propose a nonparametric Bayesian approach to functional clustering models using the spectral representation of the nonparametric unknown functions and a Dirichlet process prior for the basis coefficients. For posterior computation, we develop a fast deterministic variational Bayesian approximation algorithm and compare it with a Markov chain Monte Carlo method based on the blocked Gibbs sampling. Simulation studies and real data application examples illustrate the performance of the proposed method. This is joint work with Seongil Jo and Taeryon Choi.

Smoothing and Mean-covariance Estimation of Functional Data with a Bayesian Hierarchical Model

Dennis Cox

Rice University, United States. *dcox@rice.edu*

Abstract: We propose a nonparametric Bayesian approach to smooth all functional observations simultaneously and non parametrically. In the proposed approach, we assume that the functional observations are independent Gaussian processes subject to a common level of measurement errors, enabling the borrowing of strength across all observations. Unlike most Gaussian process regression models that rely on pre-specified structures for the covariance kernel, we adopt a hierarchical framework by assuming a Gaussian process prior for the mean function and an Inverse-Wishart process prior for the covariance function. These prior assumptions induce an automatic mean-covariance estimation in the posterior inference in addition to the simultaneous smoothing of all observations. Such a hierarchical framework is flexible enough to incorporate functional data with different characteristics, including data measured on either common or uncommon grids, and data with either stationary or non stationary covariance structures. Simulations and real data analysis demonstrate that, in comparison with alternative methods, the proposed Bayesian approach achieves better smoothing accuracy and comparable mean-covariance estimation results. Furthermore, it can successfully retain the systematic patterns in the functional observations that are usually neglected by the existing functional data analyses based on individual-curve smoothing.

Tue, June 28 (13:30-15:10) | IP23 | Sponsor: Chinese Society of Probability and Statistics

Recent Advances in Design of Experiments

Organizer: Min-Qian Liu (Nankai University)

Chair: Chongqi Zhang (Guangzhou University)

Experimental Designs for Radar Countermeasure Reconnaissance Equipment

Yu Tang

Soochow University, China. *ytang@suda.edu.cn*

Abstract: Under complex electromagnetic environment, there are many challenges for experiments of radar countermeasure reconnaissance equipment. For example, they may contain both quantitative and qualitative factors. Meanwhile, some factors should satisfy certain constraints, which makes the experimental region irregular. In this talk, I will give some partial solutions to these problems.

Column-orthogonal Designs and Orthogonal Latin Hypercube Designs with Multi-dimensional Stratification

Jian-Feng Yang

Nankai University, China. *jfyang@nankai.edu.cn*

Abstract: In this talk, we introduce some methods to construct orthogonal Latin hypercube designs and column-orthogonal designs with multi-dimensional stratification by rotating asymmetric and symmetric orthogonal arrays. Apart from orthogonality, the resulting designs also preserve better space-filling property than those constructed by using the existing methods. In addition, we provide a method to construct a new class of orthogonal Latin hypercube designs with multi-dimensional stratification by rotating regular factorial designs. Some newly constructed orthogonal Latin hypercube designs are tabulated for practical use.

Sensitivity Analysis for Computer Experiments using Permutations

Shifeng Xiong

Chinese Academy of Sciences, China. *xiong@amss.ac.cn*

Abstract: This paper presents a permutation-based sensitivity index to measure the effect of an input on the output of a model. The proposed method is related to the statistical testing problem for testing the significance of the input, and thus possesses some frequentist properties which the current sensitivity analysis methods do not have. The implementation of this method for computer experiments is discussed in details. Numerical simulations and a real application are presented to illustrate the proposed method.

Tue, June 28 (13:30-15:10) l IP25 l Sponsor: Chinese Statistical Association (Taiwan)

Biostatistics and Biomedical Statistics

Organizer: Tsung-Chi Cheng (National Chengchi University)

Chair: Jeng-Min Chiou (Academia Sinica)

An Additive-multiplicative Hazard Regression Model with Longitudinal Covariates

Yi-Kuan Tseng

National Central University, Taiwan. *tseng.yk@gmail.com*

Abstract: In this talk, a joint model of semiparametric additive-multiplicative hazards model and longitudinal processes is proposed for specifying the relationship between survival time and longitudinal covariates. The additive-multiplicative hazards models include some popular survival models in the multiplicative part such as the Cox proportional hazards model, the accelerated failure time model and the extended hazard model, along with an additive part Aalen additive model. Under the nested structure, the additive-multiplicative hazards model may provide a statistical tool to check the adequacy of using an additive hazard regression model or a multiplicative model. A pseudo joint likelihood approach is proposed to estimate the unknown parameters and components through a Monte Carlo EM algorithm. A case study of evaluating the association between the survival time of AIDS patients and the biomarkers, CD4 counts and viral loads demonstrates the effectiveness of the procedure. This is a joint work with Tzu-Ling Wang from National Hsinchu University of Education.

Conducting Non-experimental Comparative Research for Cancer Treatments using Large-scale Database

Yi-Hsin Yang

Kaohsiung Medical University, Taiwan. *yihsya@kmu.edu.tw*

Abstract: Large-scale database has become commonly available in many health care systems. Data analysis with statistical modelling nowadays has been recognized as a useful platform for medical research in evaluating effectiveness medical treatments. The large number of collected observations has the benefit of statistical power at relatively low cost. However, not being able to prospectively follow-up and randomization may reduce the internal validity, and hence the implication of results will be limited. While the internal validity is an important issue in conducting comparative studies, many approaches have been developed to provide resolutions. In terms of deviation, one may compute propensity scores to incorporate various sources of possible dispersion. Therefore, the extent of possible variables to be included in computation is needed to be identified. In terms of statistical modelling, one could also choose from adjusting by explanatory variables or applying weights in observations. Different approaches would result in different performance in effect sizes, computation/programming resources, statistical powers as well as internal validity. In this presentation, a variety of approaches will be applied in a comparative cancer treatment study. Practical recommendations will be delivered in terms of future studies for cancer research scenario.

Robust Diagnostics for Multivariate Data with a Mixture of Continuous and Categorical Variables

Tsung-Chi Cheng

National Chengchi University, Taiwan. *chengt@nccu.edu.tw*

Abstract: TCheng and Biswas (2008) apply the maximum trimmed likelihood estimator (MTLE) to obtain the robust estimators of multivariate location and shape, especially for data mixed with continuous and categorical variables. However, the detection of outliers for this approach relies on the Mahalanobis distances based on the continuous part. This paper adapts the approach proposed by Riani, Atkinson, and Cerioli (2009) to form the importance of envelopes for each log-likelihood based on MTLE. The resulting graph enhances the structure of forward plots and identifies potential outliers by taking both continuous and discrete parts into account. A simulation study and real data analysis illustrate the performance of the proposed approach in the data of this kind.

Gene Set Correlation Analysis

Chen-An Tsai

National Taiwan University, Taiwan. *catsai@ntu.edu.tw*

Abstract: Gene set analysis (GSA) aims to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. A variety of GSA methods have been proposed to assess the enrichment of sets of genes with respect to mean differences for a categorical phenotype. However, a very limited number of methods have been proposed to address the problem in identifying enriched gene sets associated with continuous phenotypes and differential covariance structure. In this paper, we develop a novel GSA test statistic, called SDR, based on sufficient dimension reduction technique, which aims to capture sufficient information between genes and the phenotype variable. The advantages of our proposed method are to allow for identification of enriched gene sets against categorical and continuous phenotypes, as well as to assess differential covariance structure.

Tue, June 28 (13:30-15:10) | IP30 | Sponsor: Japan

Recent Advances on Algebraic Methods in Statistics

Organizer: Satoshi Aoki (Kobe University)

Chair: Satoshi Kuriki (The Institute of Statistical Mathematics)

Sampling Methods of Fractional Factorial Designs

Satoshi Aoki

Kobe University, Japan. *aoki@math.kobe-u.ac.jp*

Abstract: The application of computational commutative algebra to design of experiments is first proposed by Pistone and Wynn (1996, Biometrika), and developed by many researchers. In this field, designs are treated by design ideals, which is the set of polynomials vanishing on the design points. The algebraic description of the design ideals such as Groebner basis gives algebraic interpretation of classical concept of fractional factorial designs such as confounding and orthogonality. In this talk, we consider algebraic approach for the problem of sampling fractional factorial designs.

Markov Bases for Logit Models with Some Designs

Hisayuki Hara

Niigata University, Japan. *hara@econ.niigata-u.ac.jp*

Abstract: The structure of Markov bases for logit models with discrete covariates is known to be complicated and so it is difficult to obtain an exact list of a Markov basis for the models in general. If we know the upper bounds of the highest degree of moves in minimal Markov bases, we can easily implement exact tests with lattice bases for these models. In this talk we show that for some models we can obtain upper bounds of moves of minimal Markov bases and we provide implementation algorithms of exact test for these models by using the upper bound and lattice basis for these models.

Loading Monotonicity of Weighted Insurance Premiums, and Total Positivity Properties of Weight Functions

Donald Richards

The Pennsylvania State University, United States. *richards@stat.psu.edu*

Abstract: Recent articles have developed the loading monotonicity properties of weighted insurance premiums, in which the premium sizes increase as certain scalar "loading" parameters increase. Those articles derived the loading monotonicity properties from the total positivity of order 2 of some classical weight functions. In our work, we derive conditions under which those same weight functions are totally positive, or strictly totally positive, of order infinity. As a consequence, we deduce higher-order monotonicity and ordering properties of the weighted losses derived from some classical weighted insurance premiums based on scalar loading parameters. We show furthermore that loading monotonicity properties can be established for weighted insurance premiums constructed with multidimensional loading parameters.

The Role of Algebraic Statistics in Estimation and Modeling of Random Graphs and Networks

Sonja Petrovic

Illinois Institute of Technology, United States. *sonja.petrovic@iit.edu*

Abstract: The ubiquity of network data in the world around us does not imply that the statistical modeling and fitting techniques have been able to catch up with the demand. This talk will discuss some of the basic modeling questions that every statistician knows are fundamental, some of the recent advances toward answering them, and the challenges that remain. The specific focus of the talk will be on goodness of fit testing for random graph models.

Recent joint work with Despina Stasi and Elizabeth Gross developed a new testing framework for graphs that is based on combinatorics of hypergraphs. More broadly, the talk will summarize a few lines of research that are intimately connected to discrete mathematics and computer science, where sampling algorithms, hypergraph degree sequences, and polytopes play a crucial role in the general family of statistical models for networks called exponential random graph models.

Tue, June 28 (13:30-15:10) | IP56

Modern Statistical Methods for Complex Data

Organizer: Haipeng Shen (The University of Hong Kong)

Chair: Haipeng Shen (The University of Hong Kong)

Are Your Co-variates Fixed Constants or Random Variables? – It Matters

Lawrence David Brown

The Wharton School, The University of Pennsylvania, United States. *lbrown@wharton.upenn.edu*

Abstract: Statistical analyses via linear models and generalized linear models involve covariates. In conventional notation these are the X-values, and the observations are the Y-values. In many conventional discussions of these models and their applications the X-values are treated as fixed constants. Here is a quote from the excellent recent text by Agresti: "GLMs treat yi as random and xi as fixed. ... In practice xi is itself often random. ... In this book we condition on its observed values in conducting statistical inference [and consequently treat the xi as if they were fixed]." Is it really OK to condition on the observed values of the random covariates and treat them as if they were fixed? The short answer is that it is sometimes OK and it is sometimes not OK. A key part of the answer depends on whether the analytical linear model or GLM is an accurate representation of the stochastic nature of the data or is "mis-specified". This talk will characterize answers to this basic question and describe valid alternative forms and targets of inference for situations involving random covariates. Consequences related to this issue in linear models lead to alternative inference for the Average Treatment Effect in randomized clinical trials, to alternative forms of the popular C_p criterion for model selection, and to improved estimates and predictions in semi-supervised learning. Most of the current talk will be devoted to an exposition of the main issue – the role of random covariates in standard methodology. Some of the consequences for specific modes of application will be discussed as time permits.

Additive Models for Functional Data

Byeong Park

Seoul National University, South Korea. *bupark@stats.snu.ac.kr*

Abstract: We propose functional additive models for functional regression with a scalar response and multiple functional predictors that are additive in the functional principal components of the predictor processes. For the case of a single functional predictor, the functional principal components are uncorrelated, so that a simple application of a marginal regression technique can be applied when it is furthermore assumed that the predictor components are independent, as for example in the case of a single Gaussian predictor process. When one has multiple functional predictors this independence assumption cannot be justified and therefore the dependency of the predictor components needs to be addressed. This motivates us to propose a new smooth backfitting technique for the estimation of the additive component functions in functional additive models with multiple functional predictors. A major difficulty in developing this technique is that the eigenfunctions and therefore the functional principal components of the predictor processes, which are the arguments of the proposed additive model, are unknown and need to be estimated from the data. We investigate how this required estimation of the functional principal components affects the estimation of the additive component functions and develop a complete asymptotic theory. We also study the finite sample properties of the proposed method through a simulation study and a real data example.

A Simple and Practical Approach Towards Testing Global Restrictions on General Functions

Valentin Patilea

ENSAI, France. *valentin.patilea@gmail.com*

Abstract: We propose a simple bootstrap procedure for inference on vectors or functions in a general context that involves estimation only under the alternative, while constraints are imposed via choice of a suitable transformation of the unconstrained estimate. The procedure is quite general and applies directly to functions or derivatives defined by separable and non-separable regression models. It can be used with parametric, semi- and nonparametric estimators without modification. Potential applications include, but are not limited to, inequality inference on mean or quantile regression models where the bounds depend on the model's covariates, checking monotonicity, convexity, symmetry, homogeneity, for multivariate functions.

Bilinear Regression with Matrix Covariates in High Dimensions

Dan Yang

Rutgers University, United States. *dyang@stat.rutgers.edu*

Abstract: Traditional functional linear regression usually takes a one dimensional functional predictor as input and estimates the continuous coefficient function. Modern applications often generate two dimensional covariates, which when observed at grid points are matrices. To avoid inefficiency of the classical method involving estimation of a two dimensional coefficient function, we propose a bilinear regression model and obtain estimates via a smoothness regularization method. The proposed estimator exhibits minimax optimal property for prediction under the framework of Reproducing Kernel Hilbert Space. The merits of the method are further demonstrated by numerical experiments and an application on real imaging data.

Tue, June 28 (13:30-15:10) l TCP21

Confidence Interval in Some Models and High Dimensional Data Analysis

Organizer: Junlong Zhao (Beijing Normal University)

Chair: Chi Tim Ng (Chonnam National University)

Confidence Distribution Inferences on the Common Value in Nonparametric Model

Xuhua Liu

China Agricultural University, China. *liuxuhua@cau.edu.cn*

Abstract: Some common values may exist in nonparametric regression functions, such as the values of a step function, the first or higher order derivatives of a piecewise polynomial function. In this paper, point-wise confidence distributions, together with the combined confidence distribution for some common value in the nonparametric regression model are developed via local polynomials fit. At first, a testing method is developed to test the equality of some values based on a combined p-value. To this end, the structure of the covariance of estimators is analyzed and orthogonal decomposition is applied to get independent p-values. On the basis of this testing results, an asymptotic combined confidence distribution of the common value is developed if it exists. The referred large sample properties of the new method are proved theoretically and numerically. The good performance of the proposed method compared with some existing methods is illustrated through simulations. Finally, a real data example shows the application of the proposed method.

A New Hierarchical Classification Model with Special Consideration on the Underlying Data Generating Process

Xiaoning Wang

Renmin University of China, China. *sdwangxiaoning@ruc.edu.cn*

Abstract: In hierarchical classification problem, the classes are arranged in a hierarchical structure, typically a tree or a directed acyclic graph (DAG). It has attracted a lot of interest on various research domains such as machine learning, text mining, bioinformatics and music genre classification. Since the application usually involves in large-scale and high-dimensional data, it is very important to consider the underlying data generating process and use proper feature selection method. For this purpose, we propose a new method. Simulation and real data analyses on text documents show that the new model works efficiently with high classification accuracy. We also investigate the theoretical properties of the method.

A New Confidence Interval in the Errors-in-variables Model

Liang Yan

Beijing Institute of Technology, China. *sjzyanliang@bit.edu.cn*

Abstract: In this paper, a fiducial generalized confidence interval is developed for the slope parameter in linear errors-in-variables model when the error variance is known. Comparing with existing intervals, simulation results show that the fiducial generalized confidence intervals maintain empirical coverages much closer to the nominal level and present dramatically smaller lengths particularly in low reliability ratios. Finally, two real data examples are provided.

Conditional Expection Improved Estimation for High Dimensional SUR model

Li Zhao

Beijing Institute of Technology, China. *198211@bit.edu.cn*

Abstract: This paper focuses on estimating of regression coefficient in high dimensional seemingly unrelated regression model. When the number of equations exceeds that of the observation, there is no estimator of the regression coefficient except ordinary least square estimator. As an alternative, a two-stage improved estimator based on the conditional expectation is proposed in this paper. The new estimator is further improved through hypothesis testing for correlation between equations. Simulations show that the new estimator outperforms the ordinary least square estimator in terms of mean squared error.

Tue, June 28 (13:30-15:10) l TCP22

Numerical Computation Methods for Multivariate Distributions and Parameter Estimation

Organizer: Masahiro Kuroda (Okayama University of Science)

Chair: Yuichi Mori (Okayama University of Science)

Fast Estimation using the EM Algorithm for Gaussian Mixture Models

Masahiro Kuroda

Okayama University of Science, Japan. *kuroda@soci.ous.ac.jp*

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. 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 epsilon algorithm to speed up the convergence of the em-EM algorithm. Then we propose a stopping condition using the Aitken's method for reducing the number of iterations and computational time of the vector epsilon acceleration of the em-EM algorithm. When obtaining the best initial value, the vector epsilon 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 epsilon acceleration of the EM algorithm using a re-starting procedure given in Kuroda et al. (2015).

Approximation to the Joint Density of Eigenvalues of a Complex Wishart Matrix

Tatsuya Kuwabara

Tokyo University of Science, Japan. *1415603@ed.tus.ac.jp*

Abstract: We consider approximations to hypergeometric function with matrix arguments by the Laplace's method in complex case. The hypergeometric function appears in the joint density of eigenvalues of a complex Wishart matrix. Using Laplace approximation, we show that the joint density of the eigenvalues can be expressed with gamma density functions when population eigenvalues are infinitely dispersed. We also discuss an application of MIMO capacity.

Graphical Method on an Omunibus test for Normality

Shigekazu Nakagawa

Kurashiki University of Science and the Arts, Japan. *nakagawa@ms.kusa.ac.jp*

Abstract: We give a new graphical method to test for normality based on an approximate joint probability density function of sample skewness and kurtosis statistics. This density function is initiated by Shenton and Bowman (1977). We illustrate our method for the practical data in comparison with qqplot and Shapiro--Wilk test. We calculate powers when the alternative distributions are contaminated normal distributions. We show our method has the same power as Shapiro--Wilk test excluding the case of some parameters. As a consequence, we show that the proposal method is advantageous to make graphical interpretation in testing for normality.

Bootstrapping Distributions on the Eigenvalues of Covariance Matrix

Hiroki Hashiguchi

Tokyo University of Science, Japan. *hiro@rs.tus.ac.jp*

Abstract: In multivariate data, the original bootstrap method for estimating the covariance matrix has disadvantage that its rank may be less than the one of the sample covariance matrix. On the other hand, Bayesian bootstrap method might overcome the disadvantage since it avoids replication of sample vectors. We discuss the numerical comparison with the original and Bayesian bootstrap methods via simulation studies. In particular, we investigate the resampling distributions for the largest and smallest eigenvalues of sample covariance matrix. In principal component analysis, we apply two bootstrap methods to some information criteria that identify the equality of population eigenvalues.

The Application of the Mathematical Model for Fashions in Human Societies

Yasushi Ota

Doshisha University, Japan. *yota@mail.doshisha.ac.jp*

Abstract: When analyzing boom phenomena by a mathematical model, one of the most interesting problems is reconciling the deviation between the expected and observed values. Thus far, various mathematical models have been proposed that fit the epidemiological data of infectious diseases. Real infectious disease data are especially well-fitted to a differential equation with time delay, and this approach is used in many medical institutions. However, when applied to booms in human social phenomena such as "fashion," the model results markedly differ from the data. Mathematical modeling of such phenomena has made little progress.

In this study, boom consumers are divided into four stages; the first stage in which they have not yet consumed the commodity, the second stage in which consumption begins after the start of the boom, the third stage in which consumption stops, and the fourth stage in which consumption continues. We then propose a new mathematical model of booms in human social phenomena based on the innovator theory and a differential equation with time delay. To evaluate the model, we fitted the model solutions to real data of booms in a human social phenomenon. Since our model admits an oscillatory solution, it reduces the deviation between the expected and observed values.

Tue, June 28 (13:30-15:10) l TCP20

Statistical Analysis of Survival and Functional Data

Organizer: Xin Yuan Song (The Chinese University of Hong Kong)

Chair: Jingheng Cai (Sun Yat-sen University)

Functional Clustering of Mouse Ultrasonic Vocalization Data

Xiaoling Dou

Waseda University, Japan. *xiaoling@aoni.waseda.jp*

Abstract: Mouse ultrasonic vocalizations (USVs) are studied in various fields of science. However, background noise and varied USV patterns in observed signals make complete automatic analysis difficult. We improve a moving average method to reduce noise, define USV data as functional data by B-spline basis functions and use multiple knots to define breakpoints for discontinuous USV calls. Finally, we classify the USV functional data with same number of breakpoints by functional clustering.

The proposed methods are shown work well for non-harmonic mouse USVs taken from laboratory mice.

Bayesian Joint Modeling for Survival Data with Latent Variables

Deng Pan

Huazhong University of Science and Technology, China. *pand.whu@gmail.com*

Abstract: We propose a cox model with latent variables to investigate the observed and latent risk factors of the failure time of interest. Each latent risk factor is characterized by correlated observed variables through an exploratory factor analysis model. We develop a Bayesian approach to analyze the proposed model. We propose a Markov chain Monte Carlo sampling scheme to obtain Bayesian estimates and their standard error estimates. Simulation studies demonstrate that the proposed method performs satisfactorily. Our model is applied to a study concerning the risk factors of chronic kidney disease for Type 2 diabetic patients.

Bayesian Adaptive Lasso for Additive Hazard Model in Current Status Data

Chunjie Wang

Changchun University of Technology, China. *cjwang2014@126.com*

Abstract: Variable selection for high dimensional data has recently received a great deal of attention. Due to the complex structure of the likelihood function, only limited developments have been made for time-to-event data where censoring is present. In this paper, we investigate variable selection problem for additive hazards models with current status data. We develop a Bayesian Adaptive Lasso to conduct simultaneous estimation and variable selection. To implement our methodology, we develop an efficient Markov chain Monte Carlo algorithm to carry out the Bayesian Adaptive Lasso procedure. Nice features including the empirical performance of the proposed methodology are demonstrated by simulation studies. The proposed method is applied to a real-life application.

Transformation Model for Sparse Functional Data

Guochang Wang

College of Economics, China. *wanggc023@amss.ac.cn*

Abstract: The most popular method to model the relationship between the scalar response and the functional predictor is the functional linear model because of its simplicity and easy interpretation. However, the linear form limits its application to data with more complicated structures and the nonparametric model suffers its low convergence rate. To combine the advantages both of the liner model and nonparametric model, in the present paper, we consider the functional transformation model (FTM). Owing to the involvement of an inverse operator, it is well known that mixed data canonical correlation analysis (MDCCA) and the method based on MDCCA are numerical instability and strong sensitive to the selection of smoothing parameters. These problems are exacerbated when the functional data is longitudinal data, on which we focus in the present paper. To avoid these disadvantages, we propose an innovative procedure called mixed data singular component analysis (MDSCA) to estimate the transformation function and the functional regression parameter, simultaneously. Otherwise, we also apply MDSCA to measure the correlation between a multivariate sample and a set of functional data. Furthermore, we will give three parts of theoretical results including the asymptotic properties of MDSCA, the correlation coefficient between the sparse functional data and multivariate and FTM, respectively. Lastly, real and simulation data examples are further presented to demonstrate the value of this approach.

Tue, June 28 (13:30-15:10) | CP04

Machine Learning Session 4

Chair: Phillip Sheung Chi Yam (The Chinese University of Hong Kong)

Robust Regression under Heterogeneous Contamination

Hironori Fujisawa

The Institute of Statistical Mathematics, Japan. *fujisawa@ism.ac.jp*

Abstract: The estimator can be defined as the minimizer of the divergence between the estimated density (e.g. empirically estimated density) and parametric density. It is known that the gamma-divergence is very useful for robust estimation against heavy contamination and can show a sufficiently small latent bias even if the contamination rate is not small. We extend this result to the regression problem. The homogeneous contamination in the regression problem is easily treated because it is very similar to the iid case. The heterogeneous contamination is not easy to be treated. We suppose that the parametric model belongs to a location-scale family and the regression model is assumed on the location parameter and then we obtain a favorable robustness even in the case of heterogeneous contamination.

A Matrix-intensive Formulation of Factor Analysis with Specific Factors Dissociated From Errors

Kohei Adachi

Osaka University, Japan. *adachi@hus.osaka-u.ac.jp*

Abstract: In the classic well-known factor analysis (FA) model the specific factors and errors are commonly put together as an error term. This contradicts to the original concept of FA, where specific factors refer to the ones, each of which explains the specific variations of a single variable. They are distinguished from the common factors, which explain the variability of all input variables simultaneously. In this paper, we consider a FA model with specific factors explicitly separated from errors. The corresponding least squares FA procedure is referred to as comprehensive FA (CompFA). We consider CompFA as a matrix-intensive approach: all model unknowns, i.e., common and specific factors, loadings, and variances of specific factors, are treated as the elements in fixed parameter matrices. The goal of the paper is to study some linear-algebraic properties of the CompFA solutions. This includes: [A] the identifiability of the model part of CompFA, [B] the identifiability of the covariances among variables, factors, and errors, [C] the indeterminacy of common and specific factor scores, and [D] the relationships to the solutions of principal component analysis.

This is a joint work with Nickolay T. Trendafilov.

Orthogonal Non-negative Matrix Tri-factorization Based on the Tweedie Family

Hiroyasu Abe

Doshisha University, Japan. *eio1001@mail4.doshisha.ac.jp*

Abstract: Orthogonal non-negative matrix tri-factorization (ONMTF), a multivariate analysis technique for approximating a given non-negative data matrix using the product of three non-negative factor matrices, has been adopted in many fields. In ONMTF, the left-side and right-side factor matrices are column orthogonal. However, these factor matrices are iteratively updated without maintaining orthogonality (while maintaining non-negativity) in many estimation algorithms. Moreover, the objective functions are not monotonically non-increasing in these algorithms. In this paper, we propose new monotonically non-increasing estimation algorithms for ONMTF while maintaining orthogonality and non-negativity of the factor matrices. Our simulation study and an application involving some document-term matrices show that our algorithm can perform better than previous algorithms in terms of clustering accuracy and factor matrix estimation accuracy. In addition, we show that our algorithm can be generalized as an algorithm based on Poisson and compound Poisson distributions, both of which belong to the Tweedie family. An update equation for the center factor matrix in the ONMTF algorithm based on a compound Poisson distribution is derived by a new auxiliary function method using an inequality of a bivariate concave function. Moreover, a second simulation study demonstrates the robustness of ONMTF based on a compound Poisson distribution.

A Factor-adjusted Multiple Testing Procedure with Application to Mutual Fund Selection

Lilun Du

The Hong Kong University of Science and Technology, Hong Kong. *dulilun@ust.hk*

Abstract: In this paper, we propose a Factor-Adjusted multiple Testing (FAT) procedure based on factor-adjusted p-values in a linear factor model involving some observable and unobservable factors, for the purpose of selecting skilled funds in empirical finance. The factor-adjusted p-values were obtained after extracting the unknown latent factors by the method of principal component. Under some mild conditions, the false discovery proportion (FDP) can be consistently estimated even if the idiosyncratic errors are allowed to be weakly correlated across units. Furthermore, by appropriately setting a sequence of threshold values approaching zero, the proposed FAT procedure enjoys model selection consistency. Both extensive simulation studies and a real data analysis on how to select skilled funds in US financial market are presented to illustrate the practical utility of the proposed method.

Tue, June 28 (15:30-17:10) l DL06 l Sponsor: ICSA

Model Selection in High-Dimensional Regression and Graphical Models and Its Applications

Chair: Jin-Ting Zhang (National University of Singapore)

Distinguished Lecturer: Tze Leung Lai (Stanford University)

A New Approach to Test-based Variable Selection

Tze Leung Lai

Stanford University, United States. *lait@stanford.edu*

Abstract: Test-based variable selection, as exemplified by partial F-tests in forward stepwise or backward elimination procedures in linear regression and their extensions to generalized linear and Cox regression models, preceded information-criterion selection and prediction-based selection procedures, and is still popular in software packages. After a brief review of these variable selection methodologies, highlighting their differences and similarities, we focus on applications in which test-based variable selection fits nicely into the goal of the study but falls short of giving a valid overall test. In this connection, we also review recent developments in post-selection inference, which has become an active area of research. We then describe a new approach to test-based variable selection that yields a valid overall test, and illustrate its applications to fault diagnosis in multi-stage manufacturing processes.

Model Selection for High-dimensional Time Series

Ching-Kang Ing

Academia Sinica, Taiwan. *cking@stat.sinica.edu.tw*

Abstract: Model selection for high-dimensional regression models has been one of the most vibrant research topics in statistics and probability in the past decade. However, most of the attention has been devoted to situations where observations are independent, and hence time series data are precluded. In this talk, I shall address model selection problems for some high-dimensional time series models, including high-dimensional stochastic regression models and high-dimensional regression models with correlated errors. I shall present rates of convergence of the orthogonal greedy algorithm (OGA) under various sparsity conditions. I shall also show that when the high-dimensional information criterion (HDIC) of Ing and Lai (2011) is used in conjunction with the OGA, the resultant predictor achieves the optimal error rate.

Positive Definiteness of High Dimensional Regularized Covariance Matrix Estimator

Johan Lim

Seoul National University, South Korea. *johanlim@snu.ac.kr*

Abstract: We study the positive definiteness (PDness) problem in covariance matrix estimation. For high dimensional data, the common sample covariance matrix performs poorly in estimating the true matrix. Recently, as an alternative to the sample covariance matrix, many regularized estimators are proposed under structural assumptions on the true covariance matrix including sparsity. They are shown to be asymptotically consistent and rate-optimal in estimating the true covariance matrix and its structure. However, many of them do not take into account the PDness of the estimator and produce a non-PD estimate. To achieve the PDness, additional regularizations (or constraints) are considered on eigenvalues, which make both the asymptotic analysis and computation much harder. In this paper, we propose a simple one-step procedure to update the regularized covariance matrix estimator which is not necessarily PD in finite sample. We show that the proposed one-step modification preserves the asymptotic properties of the initial regularized estimator if the shrinkage parameters are carefully selected. In addition, it is optimization-free and has advantages in computation over existing optimization based sparse PD estimators. We apply the proposal to two multivariate procedures relying on the covariance matrix estimator - the linear minimax classification problem and the well-known mean-variance portfolio optimization problem – and show that it substantially improves the performances of both procedures.

Tue, June 28 (15:30-17:10) | IP06 | Sponsor: IMS

Recent Advances in Big Data Inference

Organizer: Yingying Fan (University of Southern California)

Chair: Daoji Li (University of Central Florida)

Trade-offs in Statistical Learning

Quentin Berthet

University of Cambridge, United Kingdom. *q.berthet@statslab.cam.ac.uk*

Abstract: I will explore the notion of constraints on learning procedures, and discuss the impact that they can have on statistical precision. This is inspired by real-life concerns such as limits on time for computation, on a budget to obtain data, or communication between agents. I will show how these constraints can be shown to have a concrete cost on the statistical performance of these procedures, and talk about management of these trade-offs.

Optimal Correlation Detection with Application to Colocalization Analysis in Dual-channel Florescence Microscopic Imaging

Ming Yuan

University of Wisconsin-Madison, United States. *ming.mingyuan@gmail.com*

Abstract: Motivated by the problem of colocalization analysis of fluorescence microscopic imaging, we study in this paper structured detection of correlated regions between two random processes observed on a common domain. We introduce a size-based normalization for the likelihood ratio statistics and show that scanning with this normalized statistics leads to optimal correlation detection over a large collection of structured correlation detection problems.

Robust Covariance Matrix Estimation via Matrix Depth

Zhao Ren

University of Pittsburgh, United States. *zren@pitt.edu*

Abstract: Covariance matrix estimation is one of the most important problems in statistics. To accommodate the complexity of modern datasets, it is desired to have estimation procedures that not only can incorporate the structural assumptions of covariance matrices, but arealso robust to outliers from arbitrary sources. In this paper, we define a new concept called matrix depth and we propose a robust covariance matrix estimator by maximizing the empirical depth function. The proposed estimator is shown to achieve minimax optimal rate under Huber's ε-contamination model for estimating covariance/scatter matrices with various structures including bandedness and sparsity.

Neyman-Pearson (NP) Classification Algorithms and NP Receiver Operating Characteristic (NP-ROC) Curves

Xin Tong

University of Southern California, United States. *xint@marshall.usc.edu*

Abstract: In many binary classification applications such as disease diagnosis, type I errors are often more important than type II errors, and practitioners have the great need to control type I errors under a desired threshold α. However, common practices that tune empirical type I errors to α often lead to classifiers with type I errors much larger than α. In statistical learning theory, the Neyman-Pearson (NP) binary classification paradigm installs a type I error constraint under some user specified level α before it minimizes type II errors. Despite recent theoretical advances, the NP paradigm has not been implemented for many classification scenarios in practice. In this work, we propose an umbrella algorithm that adapts popular classification methods, including logistic regression, support vector machines, and random forests, to the NP paradigm. Powered by these NP classification methods, we propose the NP receiver operating characteristic (NP-ROC) curves, a variant of the receiver operating characteristic (ROC) curves, which have been widely used for evaluating the overall performance of binary classification methods at all possible type I error thresholds. Despite conceptual simplicity and wide applicability, ROC curves provide no reliable information on how to choose classifiers whose type I errors are under a desired threshold with high probability. In contrast, NP-ROC curves serve as effective tools to evaluate, compare and select binary classifiers with prioritized type I errors. We demonstrate the use and advantages of NP-ROC curves via simulation and real data case studies.

Tue, June 28 (15:30-17:10) l IP27 l Sponsor: Hong Kong

Time Series and Econometrics

Organizer: Guodong Li and Philip L.H. Yu (The University of Hong Kong)

Chair: Guodong Li (The University of Hong Kong)

Testing for Stability of the Mean of Heteroskedastic Time Series

Liudas Giraitis

Queen Mary University of London, United Kingdom. *L.Giraitis@qmul.ac.uk*

Abstract: Time series models are often fitted to the data without preliminary checks for stability of the mean and variance, conditions that may not hold in much economic and financial data, particularly over long periods. Ignoring such shifts may result in fitting models with spurious dynamics that lead to unsupported and controversial conclusions about time dependence, causality, and the effects of unanticipated shocks. In spite of what may seem as obvious differences between a time series of independent variates with changing variance and a stationary conditionally heteroskedastic (GARCH) process, such processes may be hard to distinguish in applied work using basic time series diagnostic tools. We develop and study some practical and easily implemented statistical procedures to test the mean and variance stability of uncorrelated and serially dependent time series. Application of the new methods to analyze the volatility properties of stock market returns leads to some unexpected surprising findings concerning the advantages of modeling time varying changes in unconditional variance.

Linear Double Autoregressive Time Series Model and its Conditional Quantile Inference

Qianqian Zhu

The University of Hong Kong, Hong Kong. *qianqzhu@hku.hk*

Abstract: This paper proposes a new conditional heteroscedastic model, called the linear double autoregressive (AR) model. Its conditional quantile inference tools are studied without imposing any moment condition on the process or the innovations, in contrast to existing inference tools for conditional heteroscedastic models which all require that the innovations have a finite variance. The existence of strictly stationary solutions to the linear double AR model is discussed, and a necessary and sufficient condition is established by borrowing the linearity of the random coefficient AR model.

We introduce the doubly weighted conditional quantile estimation (CQE) for the model, where the first set of weights ensures the asymptotic normality of the estimators, and the second improves its efficiency through balancing individual CQEs across multiple quantile levels. Finally, goodness-of-fit tests based on the quantile autocorrelation function are suggested for the adequacy of fitted models. Simulation studies indicate that the proposed inference tools perform well in finite samples, and an empirical example is presented to illustrate the usefulness of the new model.

.

COBra: Copula-based Portfolio Optimization

Marc Paolella

University of Zurich, Switzerland. *marc.paolella@bf.uzh.ch*

Abstract: A t-copula with noncentral Student's t APARCH margins is used as a model for asset returns. Using a non-standard method of estimation, the model can be estimated nearly instantaneously, and its performance enhanced using shrinkage methods. The expected shortfall of the portfolio distribution is obtained via a combination of simulation and parametric approximation for speed enhancement. An approximate method for mean-expected shortfall portfolio optimization based on simulation is presented. Performance of the model is compared to common benchmarks.

Generalized Poisson Autoregressive Models for Time Series of Counts

Cathy W.S. Chen

Feng Chia University, Taiwan. *chenws@mail.fcu.edu.tw*

Abstract: To better describe the characteristics of time series of counts such as over-dispersion, asymmetry, structural change, and a large proportion of zeros, this paper considers a class of generalized Poisson autoregressive models that properly capture flexible asymmetric and nonlinear responses through a switching mechanism. We also investigate zero-inflated generalized Poisson autoregressive models with a structural break that can cope with data having a large portion of zeros and changes in dynamics. We employ an adaptive Markov Chain Monte Carlo (MCMC) sampling scheme to locate the structural break and to estimate model parameters. As an illustration, we conduct a simulation study and empirical analysis of New South Wales crime data sets. Our findings show a remarkable improvement by modeling the data based on such generalized Poisson autoregressive models and the Bayesian method.

Tue, June 28 (15:30-17:10) | IP29 | Sponsor: Local Host

Education of Data Science and Statistics

Organizer: Helen Meng (The Chinese University of Hong Kong)

Chair: Qi-Man Shao (The Chinese University of Hong Kong)

The Avoidance of Data Contamination in Big Data Collection Processes

C.K. Wong

iASPEC Technologies Group, Hong Kong. *ckwong@iaspec.com*

Abstract: Statistical methods can be effectively used in filtering out unwanted noises in data. More fundamentally, we also need to strengthen the ability to prevent falsified and erroneous data from entering into the system from the onset. Authentication of the data sources and traceability of its origins in order to filter out intended or unintended contaminants is one of the problems that some of the Data Scientists are actively solving.

Educating the Next Generation of Data Scientists

Helen Meng

The Chinese University of Hong Kong, Hong Kong. *hmmeng@se.cuhk.edu.hk*

Abstract: The data scientist position has been hailed "the sexiest job of the 21st century". Data scientists are needed in many different fields but talents are in short supply. This panel discussion explores how we should design an educational program to train the next generation of data scientists. We will consider the fundamental grounding needed, including mathematics and statistics, computer science, as well as application domain expertise., etc. Panel members from academia and industry will contribute their perspectives and views. The panel will also include an interactive session where the audience may contribute their ideas, suggestions and insights.

Recent Advances in Transfer Learning

Qiang Yang

The Hong Kong University of Science and Technology, Hong Kong. *qyang@cse.ust.hk*

Abstract: We often encounter situations where we have an insufficient amount of high-quality data in a target domain, but we may have plenty of auxiliary data in related domains. Transfer learning aims to exploit these additional data to improve the learning performance in the target domain. In this talk, I will give an overview on some recent advances in transfer learning for challenging data mining problems. I will present structural transfer-learning solutions under heterogeneous feature representations. I will also survey cross-domain transfer learning solutions in online recommendation, social media and social network mining. I will discuss some current limitations of cross-domain transfer learning and explore possible future directions.

From Big Data to Precision Medicine: The role of Statisticians

Feifang Hu

The George Washington University, United States. *feifang@gwu.edu*

Abstract: Precision medicine (PM) (also called personalized medicine) is a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products being tailored to the individual patient. To develop precision medicine, we need new approaches to drug-development as following: (i) Collect more genomic Big Data directly from patients; (ii) Identify important biomarkers (genes) that seem to be linked with diseases based on Big Data; (iii) Develop precision medicine based on well designed clinical trials. In this talk, I will discuss the role of statisticians in each of the procedures. Then, I will propose some new adaptive designs (of clinical trials) to deal with the special features of precision medicine.

Tue, June 28 (15:30-17:10) | IP24 | Sponsor: Chinese Statistical Association (Taiwan)

Recent Advances in Functional Data Analysis

Organizer: Jeng-Min Chiou (Academia Sinica)

Chair: Yi-Kuan Tseng (National Central University)

Modeling Functional Data Vectors

Jeng-Min Chiou

Academia Sinica, Taiwan. *jmchiou@stat.sinica.edu.tw*

Abstract: Functional data vectors consist of samples of multivariate data in which each component is a random function. We introduce a novel pairwise interaction model that leads to an interpretable decomposition of multivariate functional data and their variation into component-specific and pairwise interaction processes. The interaction processes quantify the degree of pairwise interactions between the components of the functional data vectors while the component-specific processes reflect the functional variation of a particular functional vector component that cannot be explained by the other components. We present the consistency results for the proposed methods and illustrate the model by applying it to sparsely sampled longitudinal data.

Sensible Functional Linear Discriminant Analysis

Ci-Ren Jiang

Academia Sinica, Taiwan. *cirenjiang@webmail.stat.sinica.edu.tw*

Abstract: The focus of this talk is to extend Fisher's linear discriminant analysis (LDA) to both densely recorded functional data and sparsely observed longitudinal data for general c-category classification problems. We propose an efficient approach to identify the optimal LDA projections in addition to managing the nonivertibility issue of the covariance operator emerging from this extension. We study the asymptotic properties of the proposed estimators and show that the asymptotically perfect classification can be achieved under certain circumstances. The performance of this new approach is further demonstrated with numerical examples.

Asymptotic Perfect Discrimination of Functional Data by Penalized Discriminant Analysis

Lu-Hung Chen

National Chung Hsing University, Taiwan. *luhung@nchu.edu.tw*

Abstract: Classification of functional data is widely used in different applications; recently, lots of works have been proposed. In particular, Delaigle and Hall (2012) showed in the binary case that either PCA or PLS achieves theoretically optimal classification rate for functional data under mild conditions. They also showed that perfect discrimination is attachable by PLS if the difference between mean functions is represented by the eigenfunctions corresponding to small eigenvalues of the common covariance function. In this work, we generalize the results of Delaigle and Hall (2012) to multiclass classification problems. We also show that classical penalized discriminant analysis (PDA, Hastie et al., 1995) achieves asymptotic perfect classification for both binary and multiclass classification problems with careful selection of the penalization parameter under certain conditions that are similar to those in Delaigle and Hall (2012).

Supervised Regularized Principal Component Analysis

Haipeng Shen

The University of Hong Kong, Hong Kong. *shenhaipeng@gmail.com*

Abstract: This talk introduces a dimension reduction framework called the supervised principal component analysis. The research is motivated by applications where the low rank structure of the data of interest is potentially driven by auxiliary variables measured on the same set of samples. The proposed method can make use of the information in the additional data to accurately extract underlying structures that are more interpretable. The model is formulated in a hierarchical fashion using latent variables, and includes the standard principal component analysis model and the reduced rank regression model as extreme cases. The asymptotic properties of parameter estimation are derived. We also extend the framework to accommodate special features of data such as high dimensionality and smoothness through regularization. Efficient parameter estimation and data-driven procedures for tuning selection are developed. Applications to bioinformatics and business analytics problems demonstrate the advantage of the proposed methodology.

Tue, June 28 (15:30-17:10) | IP31 | Sponsor: Japan

Recent Advances in Non- and Semi-Parametric Inference

Organizer: Yoshihiko Maesono (Kyushu University)

Chair: Toshio Honda (Hitotsubashi University)

Statistical Estimation of Composite Risk Functionals and Risk Optimization Problems

Spiridon Ivanov Penev

The University of New South Wales, Australia. *s.penev@unsw.edu.au*

Abstract: We address the estimation of composite functionals which may be nonlinear in the probability measure. We are motivated by the need to estimate coherent risk measures, which are increasingly popular in finance, insurance, and other areas associated with optimization under uncertainty and risk. We establish central limit formulae for composite risk functionals. Furthermore, we discuss the asymptotic behaviour of optimization problems whose objectives are composite risk functionals and we establish asymptotic results on their optimal values when an estimator of the risk functional is used.

Furthermore, we analyse the behaviour of the maximum Sharpe ratio as a risk indicator in portfolio allocation. We focus on non-normal returns and on long-only portfolios when a restricted optimization needs to be solved to find the estimator. The main question of interest is whether or not to include additional asset classes or return drivers in the portfolio. This corresponds to testing about significant increase in the maximum Sharpe ratio. A test is formulated without assuming multivariate normal return distributions and unlimited short-selling. Interestingly, we are able to formulate a criterion of whether or not a local solution to the restricted optimization problem is indeed a global solution.

Some Boundary-bias-free Density Estimators

Yoshihide Kakizawa

Hokkaido University, Japan. *kakizawa@econ.hokudai.ac.jp*

Abstract: We consider estimation of the probability density for nonnegative data. In that case, the standard kernel density estimator is, in general, inconsistent near the boundary, due to the so-called boundary bias (see, e.g., Wand and Jones (1995)). Many authors have suggested some remedies for removing the boundary bias. Jones (1993) gave an extensive review of the boundary corrections (re-normalization, reflection, and generalized jackknifing) until 1993. On the other hand, over the last decade, there has been growing interest in the use of asymmetric kernel (AK), whose support matches the support of the density to be estimated. Such an approach for avoiding the boundary bias is the main focus of this talk.

To the best of our knowledge, Siliverman (1986) first mentioned the idea of choosing gamma or log-normal (LN) kernel, in his book. Some specific proposals of AK density estimators and their asymptotic analyses date back to Chen (1999). For the nonnegative data, Chen (2000) first suggested the boundary-bias-free nonparametric density estimator using gamma kernel. This type of research since Chen (1999,2000) has received considerable attention as "nonparametric AK estimation" in the literature, e.g., Jin and Kawczak (2003) used Birnbaum-Saunders (BS)/LN kernel, Scaillet (2004) inverse Gaussian (IG)/reciprocal inverse Gaussian (RIG) kernel, Koul and Song (2013) inverse gamma kernel, Marchant et al. (2013) (see also Saulo et al. (2013)) a subfamily of generalized BS (GBS) kernels, Igarashi and Kakizawa (2014) a family of generalized inverse Gaussian kernels, Hirukawa and Sakudo (2015) a family of generalized gamma kernels, and Igarashi (2015) a family of weighted LN kernels.

We discuss some AK density estimators, because of the "bad" formulation of the above-mentioned IG, RIG, LN, and (G) BS kernel density estimators. This talk is based on Igarashi and Kakizawa (2014) and Igarashi (2015), including some recent developments in this area.

Testing Symmetry of Unknown Densities via Smoothing with the Generalized Gamma Kernels

Masayuki Hirukawa

Setsunan University, Japan. *hirukawa@econ.setsunan.ac.jp*

Abstract: This paper improves the test of symmetry by Fernandes, Mendes and Scaillet (2015) through combining it with the generalized gamma kernels, a new class of asymmetric kernels proposed by Hirukawa and Sakudo (2015). It is demonstrated that the improved test statistic has a normal limit under the null of symmetry and is consistent under the alternative. A test-oriented smoothing parameter selection method is also proposed to implement the test. Monte Carlo simulations indicate superior finite-sample performance of the test statistic. It is worth emphasizing that the performance is grounded on the first-order normal limit and a small number of observations, despite a nonparametric convergence rate and a sample-splitting procedure of the test.

Asymptotic Properties of Kernel Type Estimators of Ratios

Taku Moriyama

Kyushu University, Japan. *moritaku3542168@gmail.com*

Abstract: In this talk, we consider direct kernel type estimators for a hazard ratio, a conditional probability and a regression function. Cwik and Mielniczuk (1989) have proposed a direct kernel type estimator of a ratio of density functions. Generalizing their idea, we introduce new direct estimators and obtain asymptotic mean squared errors. Hazard ratio has a fundamental role in survival analysis and conditional density is also important in multivariate analysis. Regression function is a conditional expectation and so we can obtain a direct kernel type estimator of the regression function, which is different from the Nadaraya-Watson estimator. The asymptotic mean squared errors of the new estimators are obtained and we compare these new estimators and the ordinal indirect estimators, both theoretically and numerically.

Tue, June 28 (15:30-17:10) | TCP24

Recent Topics in Medical Statistics

Organizer: Shu-Hui Chang (National Taiwan University)

Chair: Shu-Hui Chang (National Taiwan University)

Confidence Intervals for the Difference Between Two Median Survival Times for Clustered Survival Data

Yu-Mei Chang

Tunghai University, Taiwan. *yumei0115@thu.edu.tw*

Abstract: Clustered survival data arise often in clinical trial design, where the correlated subunits from the same cluster are randomized to different treatment groups. Under such design, we consider the problem of constructing confidence interval for the difference of two median survival time given the covariates. We use Cox gamma frailty model to account for the within-cluster correlation. Based on the conditional confidence intervals, we can identify the possible range of covariates over which the two groups would provide different median survival times. The associated coverage probability and the expected length of the proposed interval are investigated via a simulation study. The implementation of the confidence intervals is illustrated using a real data set.

Statistical Inference on Censored Data for Targeted Clinical Trials under Enrichment Design

Chen-Fang Chen

National Taiwan University, Taiwan. *d96621203@ntu.edu.tw*

Abstract: For the traditional clinical trials, inclusion and exclusion criteria are usually based on some clinical endpoints, the genetic or genomic variability of the trial participants are not totally utilized in the criteria. After completion of the human genome project, the disease targets at the molecular level can be identified and can be utilized for the treatment of diseases. However, the accuracy of diagnostic devices for identification of such molecular targets is usually not perfect. Some of the patients enrolled in targeted clinical trials with a positive result for molecular target might not have the specific molecular targets. As a result, the treatment effect may be underestimated in the patient population truly with the molecular target. To resolve this issue, under the exponential distribution, we develop inferential procedures for the treatment effects of the targeted drug based on the censored endpoints in the patients truly with the molecular targets. Under an enrichment design, we propose using the EM algorithm in conjunction with the bootstrap technique to incorporate the inaccuracy of the diagnostic device for detection of the molecular targets on the inference of the treatment effects. A simulation study was conducted to empirically investigate the performance of the proposed methods. Simulation results demonstrate that the proposed estimator is unbiased with adequate precision and the confidence interval can provide adequate coverage probability. In addition, the proposed testing procedure can adequately control the size with sufficient power. A numerical example illustrates the proposed procedures.

Dynamic Survival Prediction Using Marker Processes

Deng-Huang Su

Far-Eastern Polyclinics, Taiwan. *dhsu888@yahoo.com.tw*

Abstract: In clinical practice, the records of patients with chronic diseases is a form of the follow-up data. At each patient’s visit, the physician will collect the signs or event information to understand the level of the patient's future risk of complications or death. According to the level of these risks, physicians need to take some appropriate actions to prevent or delay the occurrence of complications or death. So, how to quantify such risks is a clinically important issue. The purpose of this paper is to use the dynamic messages of marker and the patients’ basic characteristics to predict the patients’ survival. Cox’s model is a very common regression model in the survival analysis and takes the advantage of long-term data with chronological features. Time-dependent Cox’s model was used to clearly construct the correlation between basic covariate, marker process and the termination event without the assumptions of the specific distributions of these variates. The problems of separately modeling the relationship of covariates between the marker process and termination event are resolved. However, the drawback is that Cox’s model is an immediate model of explanation and could not directly be used in the prediction of the survival. It is needed to estimate the basic hazard function and determine the course of marker for the purpose of prediction. Therefore, we used Bayes' theorem and conditional probability to overcome these problems. We could use the conditional distribution of basic covariates and marker process, giving appropriate weights, to estimate the future probability of survival under different conditions. The advantage of this method is that it can avoid the estimation of basic hazard function of the Cox’s model and the marginal distribution of marker process. The method is illustrated with the example of papillary thyroid carcinoma.

An Adaptive Procedure to Construct Robust Genetic Association Tests using Case-parents Triad Family Data

Jiun-Yi Wang

Asia University (Taiwan), Taiwan. *jjwang@asia.edu.tw*

Abstract: A model-specific test can be powerful to test genetic associations when the underlying genetic model is correctly specified. However, it can lose substantial power when the model is misspecified. Several methods have been proposed to deal with the problem, such as the maximum test statistic, the maximum efficiency robust test and the constrained likelihood ratio test. These tests have been shown to be robust against model misspecification, but they are either time-consuming in computation or not sufficiently high in power robustness under some situations. Using case-parents triad family data, in this study we proposed a data-driven procedure to construct two adaptive robust genetic association tests. One of the tests is simple in calculation and has fair power robustness. The other test is very powerful and even comparable with those model-specific tests. The results show that the proposed adaptive procedure could be beneficial to large-scale association studies.

Tue, June 28 (15:30-17:10) l TCP27

Recent Developments in Biostatistics and Spatial Statistics

Organizer: Yingan Liu (Nanjing Forestry University)

Chair: Junhao Pan (Sun Yat-sen University)

The Weighting Bi-level Penalized Method for the Identification of Rare and Common Variants in Genetic Associations Studies

Jianwei Gou

Nanjing Forestry University, China. *gjw1983@139.com*

Abstract: Multiple rare variants within the same gene can contribute to largely monogenic disorders, so data collection in genetic association study has turned toward the exome and whole genome sequencing. However, it is well known that the single marker methods frequently used for common variants have low power to detect rare variants associated with disease. The methods that collapses the rare variants in the gene into a single variable have been shown to be efficient and powerful to identify rare variants. For example, the group penalized methods have been applied into the detection of the subset of rare variants most associated with the disease. We proposed WeGEL(Weighting Group Exponential Lasso) which incorporates a burden-based weighting of the rare variants. We demonstrate that the WeGEL has a number of statistical property over previously proposed group penalties and bi-level penalized to select the rare variants. Finally, we apply these methods into the detecting rare and common variants in a genetic association study.

Geospatial Analysis Between Green Infrastructure Distribution and Infectious Disease Datasets

Jing Shen

Nanjing Forestry University, China. *shenjing@njfu.edu.cn*

Abstract: We developed Bayesian Hierarchical modeling to understand the relationship between Green infrastructure (GI) two dimension spatial distributions and the rate of infectious disease (ID) breakout. GI changes associated with public health, GI and rate of ID are all closely related with the climate, we focused on a relationship between spatial processes involved in GI and ID datasets. The Hand-Foot-Mouth Disease (HFMD) is the most common infectious disease in children. We connected the results which the normalization of difference vegetation index for remote sensing image to the Xi'an city HFMD data surveillance. HFMD cases from the 2008–2012 periods. Based on our datasets, particularly on quantifying four types of GI: green open spaces (primarily public parks), shade trees, green roofs, and vertical greening systems (green walls and facades), calculated GI index which we can quantitative the control disease function of GI. Bayesian spatial modeling is more flexible in generalized linear modeling, is corporate with R program, we use package about Bayesian Hierarchical Model (BHM) and get the results: After controlling spatial individual effect and spatial dependence, the coefficients of GI index to rate of HFMD are given. Comparing with the time series analysis and cross-sectional analysis, BHM had more advantages in making full use of the monitoring information and analysis the spatial heterogeneity and spatial dependence of the data.

We provide a map of hot relationships sites that may be show the spatial character parameters as an example of practical application of our work. The model results were applied city within the GIS and give advices for the urban landscape planning.

Robust Inferences for Latent Variable Model Mixed with Hidden Markov Model

Yemao Xia

Nanjing Forestry University, China. *ym_xia71@163.com*

Abstract: Hidden Markov model (HMM) is a widely appreciated statistical tool to explore potential heterogeneity of data and explain dynamic patterns of hidden states across the time. However, statistical results based on HMM are often sensitive to model mis-specifications and/ourliters, consequently, not robust against the model deviations. In this paper, we extended HMM to the latent variable model mixed with the HMM and developed a robust procedure for analyzing such model. A class of normal scale mixture models is established for LVM mixed with HMM, to accommodate the heavy-tailed data or outliers. Within the frequency statistical framework, Monte Carlo expectation conditional maximization (MCECM) algorithm is implemented to carry out statistical analysis. Parameters expansion technique is adopted here to speed up the convergence of the algorithm. Moreover, some asymptotical results are established to illustrate the large sample behavior of the method. To determine the number of laten state as well as the type of the mixing distributions, deviance information criteria (DIC) and L_v measure are used to take model selection. Simulation studies and real data analysis illustrate that ignoring the effects of outliers will lead to serious biases for statistical inferences.

Block-based Association Tests for Rare Variants Using Kullback-Leibler Divergence

Degang Zhu

Nanjing Forestry University, China. *zhudegang@gmail.com*

Abstract: Although genome-wide association studies (GWAS) have successfully detected numerous associations between common variants and complex diseases, these variants typically can only explain a small part of the heritable component of a disease. With the advent of next-generation sequencing, attention has turned to rare variants. Recently, a variety of approaches for detecting associations of rare variants have been proposed, including the Kullback-Leibler distance based tests (KLTs) for detecting genotypic differences between cases and controls. However, few of these approaches consider linkage disequilibrium (LD) structure among rare variants and common variants. In this study, we propose two block-based association tests for testing the effects of rare variants on a disease. The main idea for this approach comes from the hypothesis that a region of interest may consist of two or more LD blocks such that single nucleotide variants (SNVs) within each block are correlated while SNVs in different blocks are independent or weakly correlated. Under this hypothesis, we propose two tests that are generalizations of the KLTs by taking the block structure into account. A simulation study under various scenarios shows that the proposed methods have well-controlled type I error rates and outperform some leading methods in the literature. Moreover, application to the Dallas Heart Study data demonstrates the feasibility and performance of the two proposed methods in a realistic setting.

Tue, June 28 (15:30-17:10) l TCP28

Recent Developments in Multiple Comparison Procedures

Organizer: Siu Hung Cheung (The Chinese University of Hong Kong)

Chair: Mohammad Arashi (Shahrood University of Technology)

Superiority and Non-inferiority Tests in Clinical Studies with Multiple Experimental Treatments

Siu Hung Cheung

The Chinese University of Hong Kong, Hong Kong. *shcheung@cuhk.edu.hk*

Abstract: The purpose of a non-inferiority trial is to assert the efficacy of an experimental treatment compared with a reference treatment (standard treatment) by showing that the experimental treatment retains a substantial portion of the treatment efficacy of the reference treatment. Statistical methods have been developed for non-inferiority trials with multiple experimental treatments in three-arm trials. In this research, we propose test procedures that test superiority of the experimental treatments when NI to the reference treatment is established. The advantage of the proposed test scheme is its additional ability to identify superior treatments, amid the test power to discover NI treatments remains comparable to previous testing procedures. Both single-step and stepwise procedures are derived.

Simultaneous Confidence Intervals for Several Quantiles of an Unknown Distribution

Anthony Hayter

University of Denver, United States. *Anthony.Hayter@du.edu*

Abstract: Given a sample of independent observations from an unknown continuous distribution, it is standard practice to construct a confidence interval for a specified quantile of the distribution using the binomial distribution. Furthermore, confidence bands for the unknown cumulative distribution function, such as Kolmogorov's, provide simultaneous confidence intervals for all quantiles of the distribution, which are necessarily wider than the individual confidence intervals at the same confidence level. The purpose of this talk is to show how simultaneous confidence intervals for several specified quantiles of the unknown distribution can be calculated using probabilities from a multinomial distribution. An efficient recursive algorithm is described for these calculations. An experimenter may typically be interested in several quantiles of the distribution, such as the median, quartiles, and upper and lower tail quantiles, and this methodology provides a bridge between the confidence intervals with individual confidence levels and those that can be obtained from confidence bands. Some examples of the implementation of this nonparametric methodology are provided, and some comparisons are made with some parametric approaches to the problem.

Latent Ordinal Regression Models with Applications in Multiple Comparisons

Tong-Yu Lu

China Jiliang University, China. *lutongyu@cjlu.edu.cn*

Abstract: Traditional ordinal regression models are very popular and widely used. However, these ordinal regression models have two main limitations in practical studies. First, there is a shortage of appropriate statistical methods for covariates-adjusted comparisons of several treatments with ordinal responses. Second, it is difficult to apply the penalty-based variable selection techniques developed in recent years to these ordinal regression models. Based on our proposed identification procedure, the latent ordinal regressions models can now be conveniently used to conduct covariates-adjusted multiple comparisons of treatments. Moreover, our proposed framework enables the incorporation of shrinkage variable selection techniques, such as LASSO and SCAD.

Statistical Calibration and Exact One-sided Simultaneous Tolerance Intervals for Polynomial Regression

Ping Yang

The Chinese University of Hong Kong, Hong Kong. *pyang.pinky@outlook.com*

Abstract: Statistical calibration using linear regression is a useful statistical tool having many applications. Calibration for infinitely many future y-values requires the construction of simultaneous tolerance intervals (STI's). As calibration often involves only two variables x and y and polynomial regression is probably the most frequently used model for relating y with x, construction of STI's for polynomial regression plays a key role in statistical calibration for infinitely many future y-values. The only exact STI's published in the statistical literature are provided by Mee et al. (1991) and Odeh and Mee (1990). But they are for a multiple linear regression model, in which the covariates are assumed to have no functional relationships. When applied to polynomial regression, the resultant STI's are conservative. In this paper, one-sided exact STI's have been constructed for a polynomial regression model over any given interval. The available computer program allows the exact methods developed in this paper to be implemented easily. Real examples are given for illustration.

Tue, June 28 (15:30-17:10) l TCP31

New Statistical Methods for Graphs and Networks

Organizer: Swati Chandna (University College London)

Chair: Yi Yu (University of Cambridge)

Hierarchical Models for Independence Structures of Networks

Kayvan Sadeghi

University of Cambridge, United Kingdom. *k.sadeghi@statslab.cam.ac.uk*

Abstract: We introduce a new family of network models, called hierarchical network models. These models allow to represent in an explicit manner the stochastic dependence among the edges. In particular, every member of this family can be associated with a graphical model defining conditional independence clauses among the edges of the network, called the dependence graph. Every network model with the dyadic independence assumption can be generalized to construct members of this new family. Using this new framework we generalize the Erdos-Renyi and beta models to create hierarchical Erdos-Renyi and beta models. We provide methods for parameter estimation as well as simulation studies for models with sparse dependence graphs.

Data Analysis using Curvature of Data Spaces and their Metric Cones

Kei Kobayashi

The Institute of Statistical Mathematics, Japan. *kei@ism.ac.jp*

Abstract: Data analysis using geometrical structure of support of each data distribution and its empirical graphs has been studied. In this talk, a methodology for changing the metric of those geometrical structures is proposed by introducing two scalar parameters. One parameter is for tuning the curvature of the support metric space and the other parameter is for tuning the curvature of an embedding metric cone. This changing of the metric and the curvature can enhance accuracy of statistical data analysis such as classification and clustering. We also propose a new class of the intrinsic means and the corresponding variances by using the proposed class of metrics and show some examples of their applications to real data. The CAT(0) property is used to measure the curvature since it can be applied to non-smooth sets and it can be used to estimate uniqueness of the intrinsic and extrinsic means of data samples. This work is a collaboration with Henry P. Wynn (London School of Economics).

Nonparametric Graphon Estimation with Covariates

Swati Chandna

University College London, United Kingdom. *s.chandna@ucl.ac.uk*

Abstract: We study a nonparametric framework for the analysis of networks based on a natural limit object termed a graphon. Recently there has been a growing interest in the statistical estimation of graphon and its use as an exploratory tool for network data analysis (Latouche and Robin (2013), Olhede and Wolfe (2014), Chan and Airoldi (2014)). The main idea is to use the stochastic block model (Holland, Laskey, and Leinhardt, 1983) to partition the set of nodes into subgroups called blocks to obtain a piecewise constant graphon estimate from the adjacency matrix. Olhede and Wolfe (2014) note that as a network becomes larger, it may become unreasonable to assume that a majority of its structure can be explained by a blockmodel with a fixed number of blocks, and hence propose an automatic bandwidth selection procedure for a given network, allowing their graphon estimate to work as a network histogram where blocks of edges play the role of histogram bins and block sizes that of histogram bandwidths or bin sizes. In practice, the set of network observables is often not limited to the adjacency matrix, and more commonly, additional attributes or covariates are also observed. Covariates may be continuous, discrete (ordinal or nominal) or a mixture of both. We show how nodal attributes or covariates can be incorporated in an exchangeable model for networks to yield better inference using graphon estimated via the local linear approach.

How Many Communities Are There?

Diego Franco Saldana

Dow Jones, United States. *diego@stat.columbia.edu*

Abstract: Stochastic block models and variants thereof are among the most widely used approaches to community detection for social networks and relational data. A stochastic block model partitions the nodes of a network into disjoint sets, called communities. The approach is inherently related to clustering with mixture models; and raises a similar model selection problem for the number of communities. The Bayesian information criterion (BIC) is a popular solution, however, for stochastic block models, the conditional independence assumption given the communities of the endpoints among different edges is usually violated in practice. In this regard, we propose composite likelihood BIC (CL-BIC) to select the number of communities, and we show it is robust against possible misspecifications in the underlying stochastic block model assumptions. We derive the requisite methodology and illustrate the approach using both simulated and real data.

Tue, June 28 (15:30-17:10) l CP15

Data Order Structure Session 2

Chair: Zhengxiao Wu (Singapore Management University)

Simultaneous Confidence Bands for the Distribution Function of a Finite Population in Stratified Sampling

Lijie Gu

Soochow University, China. *gulijie@suda.edu.cn*

Abstract: Stratified sampling is one of the most important survey sampling approaches and widely used in practice. In this paper, we propose estimators of the distribution function of a finite population in stratified sampling by the empirical distribution function (EDF, nonsmooth) and kernel distribution estimator (KDE, smooth), respectively. Under general conditions, the rescaled estimation error processes are shown to converge to a weighted sum of transformed Brownian bridges. Moreover, simultaneous confidence bands (SCBs) are constructed for the distribution function based on EDF and KDE. Simulation experiments and analysis of a real data show that the coverage frequencies of the proposed SCBs are close to the nominal confidence level via bootstrap technique under optimal allocation rule.

Blocking in Partially Replicated Two-level Factorial Designs

Shin-Fu Tsai

National Taiwan University, Taiwan. *shinfu@ntu.edu.tw*

Abstract: Blocking is a practical technique for experimentation which allows a researcher to quantify the experimental error precisely. In this talk, I am going to introduce a new class of blocking schemes for two-level factorial designs. A noteworthy feature of the proposed designs is that the within-block and between-block replicates are both conducted, leading to that the model independent estimates for the variance components can be obtained easily. Based on these repeated design points, formal testing procedures will be presented for identifying active factorial effects as well as significant block variance component. The proposed design and analysis methods will be illustrated through a simulated experiment, and the construction method of proposed designs will be discussed.

Some results on multi-level factorial designs in complex coding

Mitsunori Ogawa

Tokyo Metropolitan University, Japan. *mitsunori_ogawa@tmu.ac.jp*

Abstract: Complex coding gives a theoretically useful framework for multi-level fractional factorial designs. Pistone and Rogantin (JSPI, 2008) utilized complex coding to generalize the theory of indicator functions for multi-level fractional factorial designs, which is originally introduced to two-level designs by Fontana et al. (JSPI, 2000). The expansion coefficients of an indicator function with respect to orthogonal contrasts have rich information on design properties. On the other hand, many concepts regarding to some specific interactions, such as clear two-factor interactions and designs of variable resolution, have been considered for two-level designs. In this talk we generalize some known results on two-level designs to multi-level designs by using complex coding and the corresponding theory of indicator functions.

Day 3
Wed, June 29

Wed, June 29 (08:30-10:10) | DL03 | Sponsor: IMS

Fusion Learning - Fusing Inferences from Diverse Sources

Chair: Mengling Liu (NYU School of Medicine)

Distinguished Lecturer: Regina Y. Liu (Rutgers University)

Confidence Distribution - An Effective Tool for Statistical Inference and Fusion Learning

Min-ge Xie

Rutgers University, United States. *mxie@stat.rutgers.edu*

Abstract: A confidence distribution (CD) is a sample-dependent distribution function that can serve as a distribution estimate, contrasting with a point or interval estimate, of an unknown parameter. It can represent confidence intervals (regions) of all levels for the parameter. It can provide "simple and interpretable summaries of what can reasonably be learned from data", as well as meaningful answers for all questions in statistical inference. An emerging theme is "Any statistical approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of CDs, as long as it can be used to derive confidence intervals of all levels, exactly or asymptotically." The developments in CD lead to useful inference tools for statistical problems where methods with desirable properties have not been available or could not be easily obtained. In this talk, we articulate the logic behind the developments, show how CD can potentially bridge posterior probabilistic inferences in Bayesian, frequentist and fiducial (BFF) schools, and provide an overview on its applications on combining information and fusion learning.

Fusion Learning for Key Comparisons

Jan Hannig

The University of North Carolina at Chapel Hill, United States. *jan.hannig@unc.edu*

Abstract: In this paper we propose a Generalized Fiducial Inference inspired method for finding a robust consensus of several independently derived collection of confidence distributions (CDs) for a quantity of interest. The resulting fused CD is robust to the existence of potentially discrepant CDs in the collection. The method uses computationally efficient fiducial model averaging to obtain a robust consensus distribution without the need to eliminate discrepant CDs from the analysis. This work is motivated by a commonly occurring problem in interlaboratory trials, where different national laboratories all measure same unknown true value of a quantity and report their CDs. These CDs need to be fused to obtain a consensus CD for the quantity of interest. When some of the CDs appear to be discrepant, simply eliminating them from the analysis is often not an acceptable approach, particularly so in view of the fact that the true value being measured is not known and a discrepant result from a lab may be closer to the true value than the rest of the results. Additionally, eliminating one or more labs from the analysis can lead to political complications since all labs are regarded as equally competent. These considerations make the proposed method well suited for the task since no laboratory is explicitly eliminated from consideration. We report results of three simulation experiments showing that the proposed fiducial approach has better small sample properties than the currently used naive approaches. Finally, we apply the proposed method to obtain consensus CDs for gauge block calibration interlaboratory trials and measurements of Newton's constant of gravitation (G) by several laboratories.

Statistical Fusion Learning: Combining Inferences from Multiple Sources for More Powerful Findings

Regina Y. Liu

Rutgers University, United States. *rliu@stat.rutgers.edu*

Abstract: Inferences from multiple databases or studies can often be fused together to yield a more powerful overall inference than individual studies alone. Fusion learning refers to the development of such effective approaches to synergize learnings from different sources. Effective fusion learning is particularly important in this era of big data, with the trove of data nowadays collected routinely from diverse sources in all domains and at all time.

Using the tracking of aircraft landing performance as an illustrative example, we present a powerful fusion learning approach. Specifically, we apply the concepts of confidence distribution (CD) and data depth to develop a new nonparametric approach for combining inferences from multiple studies for a common hypothesis. We discuss several new approaches in fusion learning in the context of combining test results from independent studies or joint modeling of data from possibly heterogeneous sources. These approaches are completely data driven and have several desirable properties. They can also provide solutions to several problems whose solutions have been elusive thus far. Examples of simulation data and real applications will be presented.

Wed, June 29 (08:30-10:10) I IP04 I Sponsor: IMS

Statistical Analysis of Dynamic Models

Organizer: Samuel Kou (Harvard University)

Chair: Siu Hung Cheung (The Chinese University of Hong Kong)

Intrinsic Noise in Nonlinear Gene Regulation Inference

Chao Du

University of Virginia, United States. *cd2wb@virginia.edu*

Abstract: Cellular intrinsic noise plays an essential role in the regulatory interactions between genes. Although a variety of quantitative methods are used to study gene regulation system, the role of intrinsic noises has largely been overlooked. Using the Kolmogorov backward equation (master equation), we formulate a causal and mechanistic Markov model. This framework recognizes the discrete, nonlinear and stochastic natures of gene regulation and presents a more realistic description of the physical systems than many existing methods. Within this framework, we develop an associated moment-based statistical method, aiming for inferring the unknown regulatory relations. By analyzing the observed distributions of gene expression measurements from both unperturbed and perturbed steady-states of gene regulation systems, this method is able to learn valuable information concerning regulatory mechanisms. This design allows us to estimate the model parameters with a simple convex optimization algorithm. We apply this approach to a synthetic system that resembles a genetic toggle switch and demonstrate that this algorithm can recover the regulatory parameters efficiently and accurately.

Efficient Parameter Inference for Dynamic Systems

Samuel W.K. Wong

University of Florida, United States. *swkwong@stat.ufl.edu*

Abstract: In this talk, we describe some recent progress on parameter inference for sets of ordinary differential equations. Dynamic systems in science and engineering are frequently described by ODE models (often nonlinear), where the form of the equations is given but the parameters are unknown and must be estimated from data. The data observed are usually noisy and collected at discrete time intervals as the system evolves. Further, models may be overparametrized or the observed data may be inadequate, leading to the nonidentifiability of some parameters. Using a Gaussian process to emulate the dynamic system and an approximate Bayesian approach, we describe how to efficiently estimate and evaluate the identifiability of parameters. Our examples to illustrate the method are drawn from the biological sciences.

Segmentation of Change-point Models

David Siegmund

Stanford University, United States. *siegmund@stanford.edu*

Abstract: To segment a sequence of independent random variables at an unknown number of change-points, we introduce two new procedures that are based on thresholding the likelihood ratio statistic and compare them with procedures from the recent literature. We also study confidence regions for the changepoints and joint confidence regions for the change-points and the parameter values based on the likelihood ratio statistic. This is joint research with Xiao Fang and Jian Li.

Fast Analysis of Dynamic Systems via Gaussian Emulator

Samuel Kou

Harvard University, United States. *kou@stat.harvard.edu*

Abstract: Dynamic systems are used in modeling diverse behaviors in a wide variety of scientific areas. Current methods for estimating parameters in dynamic systems from noisy data are computationally intensive (for example, relying heavily on the numerical solutions of the underlying differential equations). We propose a new inference method by creating a system driven by a Gaussian process to mirror the dynamic system. Auxiliary variables are introduced to connect this Gaussian system to the real dynamic system; and a sampling scheme is introduced to minimize the ‘distance’ between these two systems iteratively. The new inference method also covers the partially observed case in which only some components of the dynamic system are observed. The method offers a substantial saving of computational time and fast convergence while still retaining high estimation accuracy. We will illustrate the method by numerical examples.

Wed, June 29 (08:30-10:10) l IP08 l Sponsor: IMS

Statistical Advances of Image Analysis and Spatial Statistics

Organizer: Zhengjun Zhang (University of Wisconsin-Madison)

Chair: Xueqin Wang (Sun Yat-Sen University)

Neuron Network Detection

Chunming Zhang

University of Wisconsin-Madison, United States. *cmzhang@stat.wisc.edu*

Abstract: In neurophysiological study, neural signals provide hidden information of interaction among neurons. The activity of a neuron's spike firing might affect the chance of another neuron to fire in a certain period of time. To study the functional connectivity, we apply some proper regularization method combined with some loss term which adapts to unspecified distributions. Simulation results show the effectiveness of the proposed method in detecting connectivities while cleaning out insignificant interactions at the same time. We apply this method to a real neurophysiological data set collected from part of the prefrontal cortex of rats during a designed experiment. The results provide some insight into the interaction network in that region.

A Multi-resolution Scheme for Analysis of Brain Connectivity Networks

Vikas Singh

University of Wisconsin-Madison, United States. *vsingh@biostat.wisc.edu*

Abstract: There is significant interest in understanding how structural/functional connectivity changes in the brain explain behavioral symptoms in neurodegenerative diseases such as Alzheimer's disease (AD). Clear variations in connectivity at the dementia stage of the disease have been identified in the literature. Despite such findings, AD-related connectivity changes in the preclinical stage of the disease still remain poorly characterized. Such preclinical datasets are typically smaller in size and group differences are subtle, making analysis challenging. This talk will describe some of our recent efforts to overcome these difficulties in an effort to elucidate how brain connectivity varies as a function of genotype and various other risk factors, even in asymptomatic individuals.

The engine driving these analyses is a new multi-resolution scheme for performing statistical analysis of connectivity networks derived from neuroimaging data. Our algorithm derives a wavelet representation at each connection edge which captures the graph context at multiple resolutions. Extensive empirical evidence shows how this framework offers improved statistical power in analyzing structural connectivity in diffusion tensor images (DTI) obtained via so-called tractography methods. We will present results showing connectivity differences between AD patients and controls that were not evident using standard approaches. Later, we will show results on individuals that are not yet diagnosed with AD but have a positive family history risk of AD where our algorithm helps in identifying potentially subtle differences between patient groups. If time permits, I will cover some very preliminary results on applying these ideas on the entire (currently available) dataset from the Human Connectome project.

Infill Asymptotics for Multivariate Spatial Processes

Hao Zhang

Purdue University, United States. *zhanghao@purdue.edu*

Abstract: Infill asymptotics provide insightful understanding of the properties of kriging. However, for cokriging, very few infill asymptotic results are available. In this talk, I will use a few examples to show that we may expect infill asymptotics are equally helpful for the understanding of cokriging and illustrate possible approaches to derive infill asymptotic results for spatial multivariate processes.

Study on Spatial Extreme Dependence Patterns in China's Smog Extreme Co-movements

Zhengjun Zhang

University of Wisconsin-Madison, United States. *zjz@stat.wisc.edu*

Abstract: China's smog problem has become very serious and has been drawing widely attentions among government administrators, ordinary civilians, and international societies. The smog does hurt not only China's economy, but also the world's economic development. It raises a great health concern, especially when extreme smog occurs. China's extreme smog outstands in its frequency, duration, severity and spatial property. This paper intends to conduct systematic statistical analysis on smog's spatial extreme dependence patterns using hourly PM2.5 data from 945 stations in China. The generalized extreme value (GEV) distribution is used to model extreme smog for each station to disclose their individual tail characteristics. Spatial maps using the inverse distance method are drawn to show its spatial diversity in tail lengths. Six smog zones are then located via estimated shape parameters. Using transformed unit Frchet scales, a class of tail quotient correlation coefficients (TQCC) is performed to investigate tail dependencies among stations within each smog zone. City families suffering simultaneous extreme smog are identified through certain algorithm on strong tail dependence pairs between and inside cities. Tail dependencies with lagged hours are also calculated to show the time difference of the extreme smog co-movements among stations as well as the spatial variety.

Wed, June 29 (08:30-10:10) I IP42

Recent Advances in Analysis of Complex Observational Data

Organizer: Kevin (Zhi) He (University of Michigan)

Chair: Yi Li (University of Michigan)

Investigating Reliability of Diagnostic Classifications using Category-specific Measures: Lessons from Multi-center Clinical Research Networks

J. Richard Landis

University of Pennsylvania, United States. *jrlandis@mail.med.upenn.edu*

Abstract: Establishing credibility in research findings frequently requires investigation of the reliability of primary measures, such as disease classifications. This talk will summarize a 3-level nested (clinic, subject, observer) category-specific variance components model to estimate agreement and disagreement patterns among multiple diagnostic classifications, potentially from unbalanced designs. These methods will be illustrated within several clinical research network studies, utilizing category-specific intraclass correlation coefficients (ICCs), together with pairwise measures of disagreement among categories, to explore the differential reliability patterns of categorical measurement scales (Landis et al 2011). In particular, for ordinal data, in which the ordered categories are constructed based on clinical diagnostic criteria, the category-specific ICCs frequently differ considerably at the extremes, relative to the ICCs for the interior categories. Moreover, reliability patterns also differ among target subgroups, which has important implications for "personalized medicine" decisions. Estimation of these category-specific agreement measures of ICCs, together with nested model-based confidence intervals, are illustrated within specialized software for several clinical research examples.

References:

Landis JR, King TS, Choi JW, Chinchilli VM, Koch GG (2011). Measures of agreement and concordance with clinical research applications. Journal of Biopharmaceutical Statistics 3(2): 185-209.

Modeling Time-varying Effect of Treatment Switching with Application to Estimating the Effect of Change in Dialysis Vascular Access

Yuedong Wang

University of California, Santa Barbara, United States. *yuedong@pstat.ucsb.edu*

Abstract: Vascular access for haemodialysis is of paramount importance. Although studies have found that central venous catheter is often associated with poor outcomes and switching to arteriovenous fistula is beneficial, it has not been fully elucidated how the effect of switching of access on outcomes changes over time and whether the effect depends on switching time. Analysis of the longitudinal data needs to account for changes over multiple time indices. We propose a flexible model that jointly models the change over treatment time, the trend over calendar time, the change associated with treatment switching and time-varying effects of covariates. The effect of switching may depend on both the time of switching and time since switching. All unknown functions are modeled nonparametrically using local linear smoothers and estimated using a backfitting procedure. Large sample properties are studied. Simulation studies show excellent finite-sample performance. Our methods are applied to investigate the effect of vascular access change on albumin in dialysis patients. It is concluded that the benefit of switching from central venous catheter to arteriovenous fistula depends on the time of switching, the sooner the better.

Big Data Regression and Prediction in Functional Genomics

Hongkai Ji

Johns Hopkins University, United States. *hji@jhu.edu*

Abstract: The rapid growth of functional genomic data makes it possible to use massive amounts of publicly available data to build models for predicting one high-throughput genomic data type from another data type. This can be formulated as a challenging big data regression problem which involves fitting millions of high-dimensional regressions simultaneously. To cope with the high dimensionality and heavy computation, we developed BIRD algorithm that leverages the correlation structure in the data to make computation fast and predictions accurate.

Nonparametric Estimation of Conditional Moments with Right-censored Selection Biased Data

Cedric Heuchenne

University of Liège, Belgium. *C.Heuchenne@ulg.ac.be*

Abstract: Suppose the random vector (X, Y) satisfies the nonparametric regression model $Y = m(X) + \varepsilon$, where $m(x) = E[Y|X = x]$ and $\sigma^2(x) = Var[\varepsilon|X = x]$ are unknown smooth functions and the error ε has zero mean and finite variance conditionally on $X = x$. The pair (X, Y) is subject to generalized selection bias while the response is possibly right-censored. We construct new estimators for $m(\cdot)$ and $\sigma^2(\cdot)$ by extending the conditional estimation methods introduced by de Un˜aA`lvarez and Iglesias-Perez, 2010. Asymptotic properties of the resulting estimators are obtained and the proposed procedures are studied via extended simulations. Finally, a data set on the mortality of diabetics is analyzed.

Wed, June 29 (08:30-10:10) | IP10 | Sponsor: ICSA

Development in Statistical Methods for the Analysis of Events

Organizer: Zhezhen Jin (Columbia University)

Chair: Wenyang Zhang (University of York)

Yakovlev Promotion Time Cure Model with Local Polynomial Estimation

Li-Shan Huang

National Tsing Hua University, Taiwan. *lhuang@stat.nthu.edu.tw*

Abstract: In modeling survival data with a cure fraction, flexible modeling of covariate effects on the probability of cure has important medical implications, which aids investigators in identifying better treatments to cure. We propose to study an extension of the Yakovlev promotion time cure model that allows for nonlinear covariate effects. We adopt the local polynomial approach and use the local likelihood criterion to derive nonlinear estimates of covariate effects on cure rates, while assuming that the baseline function in the model follows a parametric distribution. This way we adopt a flexible method to estimate the cure rate, the important part in cure models, and a convenient way to estimate the baseline function, which is less useful in practice. Asymptotic properties of local polynomial estimates are investigated. The proposed method is illustrated by simulations and real data analysis. This is a joint work with Li-Hsiang Lin.

Analysis of Stratified Mark-specific Proportional Hazards Models under Two-phase Sampling with Application to HIV Vaccine Efficacy Trials

Yanqing Sun

The University of North Carolina at Charlotte, United States. *yasun@uncc.edu*

Abstract: An objective of preventive HIV vaccine efficacy trials is to understand how immune responses to specific protein or sub-protein sequences of HIV associate with the level of vaccine efficacy to prevent infection with sequences of HIV targeted by the immune responses. The vaccine-induced immune response biomarkers are often measured via two-phase sampling for efficiency. Motivated by this objective, we investigate the stratified mark-specific proportional hazards model under two-phase sampling, where the mark is the genetic distance of an infecting HIV sequence to an HIV sequence represented inside the vaccine. The estimation and inference procedures based on inverse probability weighting of complete-cases and the augmented inverse probability weighted complete-case are developed. The asymptotic properties are derived, and their finite-sample performances are examined in a comprehensive simulation study. The methods are shown to have satisfactory performance, and are applied to the RV144 vaccine trial to assess whether immune response correlates of HIV infection are stronger for HIV infecting sequences similar to the vaccine than for sequences distant from the vaccine. This is a joint work with Guangren Yang, Jinan University, China; Li Qi, University of North Carolina at Charlotte, USA; and Peter B. Gilbert, University of Washington and Fred Hutchinson Cancer Research Center, USA.

Nonparametric Inference for Current Status Data

Xingqiu Zhao

The Hong Kong Polytechnic University, Hong Kong. *xingqiu.zhao@polyu.edu.hk*

Abstract: This paper studies self-normalized limits and moderate deviations of nonparametric maximum likelihood estimators for unknown cumulative distribution functions based on current status data. For this propose, two estimators for the density of the unknown cumulative distribution function are proposed and the consistency and exponential convergence of the proposed estimators are established. A self-normalized weak convergence theorem and a self-normalized moderate deviation principle are derived and applied to constructing pointwise confidence intervals and solving hypothesis testing problems.

Estimation of Concordance Probability with Censored Regression Models

Zhezhen Jin

Columbia University, United States. *zj7@cumc.columbia.edu*

Abstract: In this talk, evaluation and comparison of various methods often arise in medical research. The concordance probability can be used to assess the discriminatory power of censored regression models. In this talk, we present the estimation of concordance probability with various censored regression models in the analysis of right censored data.

Joint work with Xinhua Liu.

Wed, June 29 (08:30-10:10) l IP13 l Sponsor: IISA

Stochastic Interacting Systems

Organizer: Sunder Sethuraman (The University of Arizona)

Chair: Rabi Bhattacharya (The University of Arizona)

Long Range Exclusion Processes

Cédric Bernardin

University of Nice (France), France. *cbernard@unice.fr*

Abstract: I will discuss various aspects of the (symmetric and asymmetric) exclusion process when the probability of jumps have heavy tails.

Two Approximations of Coupled KPZ Equations

Tadahisa Funaki

The University of Tokyo, Japan. *funaki@ms.u-tokyo.ac.jp*

Abstract: In a joint work with Jeremy Quastel, we introduced an approximation of KPZ equation, which is suitable for studying the invariant measures, and investigated its limit due to Cole-Hopf transform. For a multi-component version of KPZ equation, a similar approximation can be introduced, but Cole-Hopf transform doesn't work. We apply paracontrolled calculus due to Gubinelli and others to study the limit. This is a joint work with Masato Hoshino.

Moderate Deviation Principles for Weakly Interacting Particle Systems

Amarjit Budhiraja

The University of North Carolina at Chapel Hill, United States. *budhiraj@email.unc.edu*

Abstract: Moderate deviation principles for empirical measure processes associated with weakly interacting Markov processes are established. Two families of models will be considered; the first corresponds to a system of interacting diffusions whereas the second describes a collection of pure jump Markov processes with a countable state space. For both cases the moderate deviation principle is formulated in terms of a large deviation principle (LDP), with an appropriate speed function, for suitably centered and normalized empirical measure processes. For the first family of models the LDP is established in the path space of an appropriate Schwartz distribution space whereas for the second family the LDP is proved in l^2 valued paths. Proofs rely on certain variational representations for exponential functionals of Brownian motions and Poisson random measures. Joint work with Ruoyu Wu.

On KPZ-Burgers Equation and Stochastic Particle Systems

Sunder Sethuraman

The University of Arizona, United States. *sethuram@math.arizona.edu*

Abstract: We discuss work on deriving types of KPZ-Burgers and other equations from scaling limits of fluctuation fields in weakly-asymmetric particle models.

Wed, June 29 (08:30-10:10) I IP55

New Developments in Statistical Genomics

Organizer: Yingying Wei (The Chinese University of Hong Kong)

Chair: Yingying Wei (The Chinese University of Hong Kong)

A Statistical Approach to Colocalizing Genetic Risk Variants in Multiple GWAS

Can Yang

Hong Kong Baptist University, Hong Kong. *eeyang@hkbu.edu.hk*

Abstract: It is widely agreed that complex human phenotypes (such as height, obesity, and psychiatric disorders) are highly polygenic and a large number of risk variants with small effects remain undiscovered. Recently, accumulating evidence suggests that many genetic variants may affect multiple seemly different phenotypes. Such a phenomenon is known as "pleiotropy". Undouble, identification of risk variants with pleiotropic effects not only helps to explain the relationship between diseases, but may also contribute to novel insights concerning the etiology of each specific disease. In this talk, we consider a statistical approach to colocalizing genetic risk variants in multiple GWAS by taking pleiotropic effects into account. An efficient algorithm was derived such that we were able to perform joint analysis of 20 GWAS within half an hour. Compared with single GWAS analysis, the statistical power of the proposed approach was improved about 15%-30% in this real data analysis. We believe that the proposed approach will greatly facilitate colocalization of genetic risk variants.

Estimating and Accounting for Tumor Purity in Methylation Microarray Analysis

Hao Wu

Emory University, United States. *hao.wu@emory.edu*

Abstract: Solid tumor sample is a mixture of cancer and normal cells. The mixing proportion (or the tumor "purity") brings extra noises and needs to be accounted for in cancer genomics data analysis. Estimating and adjusting for tumor purity have gained tremendous interests lately. Several computational methods and software tools were developed using gene expression, DNA methylation, copy number variation or point mutation data.

We discover that the methylation measurements from Illumina Infinium 450k are very informative for predicting tumor purities. We develop simple and efficient methods for tumor purity estimation, and differential methylation detection adjusting for purity. Analyses of a number of datasets from The Cancer Genome Atlas (TCGA) demonstrate improved performance of the proposed methods.

Spatial Temporal Modeling of Gene Expression Dynamics During Human Brain Development

Hongyu Zhao

Yale University, United States. *hongyu.zhao@yale.edu*

Abstract: Human neurodevelopment is a highly regulated biological process, and recent technological advances allow scientists to study the dynamic changes of neurodevelopment at the molecular level through the analysis of gene expression data from human brains. In this talk, we will focus on the analysis of data sampled from 16 brain regions in 15 time periods of neurodevelopment. We will introduce a two-step statistical inferential procedure to identify expressed and unexpressed genes and to detect differentially expressed genes between adjacent time periods. Markov Random Field (MRF) models are used to efficiently utilize the information embedded in brain region similarity and temporal dependency in our approach. We develop and implement a Monte Carlo expectation-maximization (MCEM) algorithm to estimate the model parameters. Simulation studies suggest that our approach achieves lower misclassification error and potential gain in power compared with models not incorporating spatial similarity and temporal dependency. We will also describe our methods to infer dynamic co-expression networks from these data. This is joint work with Zhixiang Lin, Stephan Sanders, Mingfeng Li, Nenad Sestan, and Matthew State.

Robust Identification of Gene-environment Interactions

Shuangge Ma

Yale University, United States. *shuangge.ma@yale.edu*

Abstract: Gene-environment interactions play an important role in disease development beyond the main effects of genetic and environmental factors. We pursue methods that can be robust to model mis-specification or contamination in response variables. Penalization is adopted for regularized estimation and marker selection. For computational feasibility, a progressive approach is developed, which can significantly reduce computer time while generating more reliable results. Simulation and data analyses have been extensively conducted.

Wed, June 29 (08:30-10:10) l TCP01

Recent Advances in Bayesian Nonparametrics

Organizer: Subhashis Ghoshal (North Carolina State University)

Chair: Jie Hu (Xiamen University)

Needles and Straw in a Haystack: Empirical Bayes Confidence for Possibly Sparse Sequences

Eduard Belitser

VU University Amsterdam, Netherlands. *e.n.belitser@vu.nl*

Abstract: In the many normal means model we construct an empirical Bayes posterior which we then use for uncertainty quantification for the unknown (possibly sparse) parameter by constructing an estimator and a confidence set around it as empirical Bayes credible ball. An important step in assessing the uncertainty is the derivation of the fact that the empirical Bayes posterior contracts to the parameter with a local (i.e., depending on the parameter) rate which is the best over certain family of local rates; therefore called oracle rate. We introduce the so called Excessive Bias Restriction under which we establish the local (oracle) confidence optimality of the empirical Bayes credible ball. Adaptive minimax results (for the estimation and posterior contraction problems) over sparsity classes follow from our local results. An extra (square root of) log factor appears in the radial rate of the confidence ball; it is not known whether this is an artifact or not.

Bayesian Variable Selection for Semi-parametric Regression Models

Weining Shen

University of California, Irvine, United States. *swn1989@gmail.com*

Abstract: We study a class of generalized partial linear models and consider selecting useful covariates for high-dimensional data within the parametric structure. We consider a Gaussian spike and slab prior for parametric component and a sieve prior for the nonparametric function estimation based on spline expansions. We show a strong selection consistency result in the sense that the posterior selects the true model with probability tending to one. The finite-sample performance has been evaluated by simulation studies.

An Online Gibbs Sampler Algorithm for Hierarchical Dirichlet Processes Prior

Yongdai Kim

Seoul National University, South Korea. *ydkim0903@gmail.com*

Abstract: The hierarchical Dirichlet process (HDP) is a Bayesian nonparametric model that provides a flexible mixed-membership to documents. In this talk, we develop a novel mini-batch online Gibbs sampler algorithm for the HDP which can be easily applied to massive and streaming data. For this purpose, a new prior process so called the generalized hierarchical Dirichlet processes (gHDP) is proposed. The gHDP is an extension of the standard HDP where some prespecified topics can be included in the top-level Dirichlet process. By analyzing various datasets, we show that the proposed mini-batch online Gibbs sampler algorithm performs significantly better than the online variational algorithm for the HDP.

(Joint work with Minwoo Chae, Byung-Yup Kang and Hyojoo Jung)

Bayesian Methods for Boundary Detection in Images

Subhashis Ghoshal

North Carolina State University, United States. *sghosal@stat.ncsu.edu*

Abstract: Boundary detection in images has many important applications. The boundary of a d-dimensional image may be viewed as a d-1 dimensional manifold, and in particular a smooth closed, not self-intersecting curve for 2D images. We consider a Bayesian approach to the problem using a prior indexed by the unit circle, or the unit sphere in higher dimension, typically constructed from a Gaussian process or a finite random series using trigonometric polynomials or spherical harmonics basis. For the most important case of 2D images, a very convenient prior is the squared exponential periodic Gaussian process. Its explicit eigen decomposition in terms of Bessel functions allows a convenient computational scheme and obtaining posterior contraction rate. We show that the posterior contracts at the minimax optimal rate and adapts to the unknown smoothness level of the curve. Simulation experiments show that the method is exceptionally efficient and robust against misspecification.

Wed, June 29 (08:30-10:10) I TCP04

Variable Selection and Applications of Semiparametric Models

Organizer: Dewei Wang (University of South Carolina)

Chair: Dewei Wang (University of South Carolina)

Embracing Blessing of Dimensionality in Factor Models

Quefeng Li

The University of North Carolina at Chapel Hill, United States. *quefeng@email.unc.edu*

Abstract: Factor model is an essential tool for exploring intrinsic dependence structures among high-dimensional random variables. Much progress has been made for estimating the covariance matrix from a high-dimensional factor model. However, blessing of dimensionality has not yet fully embraced in the literature: much of the available data is often ignored in constructing covariance matrix estimates. If our goal is to accurately estimate a covariance matrix of a set of targeted variables, shall we employ additional data, which are beyond the variables of interest, in the estimation? In this paper, we provide sufficient conditions for an affirmative answer, and further quantify its gain in terms of Fisher information and convergence rate. In fact, even an oracle-like result (as if all the factors were known) can be achieved when a sufficiently large number of variables is used. The idea of utilizing data as much as possible brings computational challenges. A divide-and-conquer algorithm is thus proposed to alleviate the computational burden, and also shown not to sacrifice any statistical accuracy in comparison with a pooled analysis. Simulation studies further confirm our advocate of the use of full data, and demonstrate the effectiveness of the above algorithm. Our proposal is applied to a microarray data example that shows empirical benefits of using more data.

Predictive Variable Selection for High-dimensional Response and Covariate

Yuan Wang

Washington State University, United States. *ywang@math.wsu.edu*

Abstract: In this paper, we aim to develop a new predictive model in the high-dimensional response-covariate setting. The problem is motivated by searching of the association between a set of genetic variables and brain imaging phenotypes. The proposed model considers a three-way decomposition of the regression coefficient with sparse structure reinforced at each level, with the covariance of the response incorporated to model the spatial correlation between neighboring imaging phenotypes. Bayesian procedures are developed for the inference on the parameters. The proposed model is applied to the study of Alzheimer's Disease with data from Alzheimer's Disease Neuroimaging Initiative (ADNI). The proposed model is use to to investigate the association between 1041 selected SNPs and 93 regions of interest inside the brain at different locations.

Quantile Regression in Survival Analysis with High Dimensional Sata

Qi Zheng

University of Louisville, United States. *qi.zheng@louisville.edu*

Abstract: Quantile regression offers a flexible platform to evaluate the heterogeneous impact of covariates upon the conditional distribution of responses and has been attracting great interests in high dimensional data analysis. In this article, we investigate the quantile regression in high dimensional survival data, which, to the best of our knowledge, has not yet been studied under commonly used conditional independent censoring by previous work. The L^1 penalties employed by the proposed method across a continuum of quantile levels not only account for the sparsities but also stochastically adjust to the accumulated estimation errors to achieve the consistent shrinkage of regression quantile estimates. We further adopt the "globally concerned" penalization framework from Zheng et al. [2015] by considering the adaptive L^1 penalties and a uniform tuning parameter to reduce the estimation bias as well as to attain the model selection consistency. We show that under mild conditions, the resulting estimator of regression coefficients enjoys both the oracle convergence rate uniformly and the weak convergence. Our method can be conveniently implemented with a simple algorithm. The theoretical and computational properties of the newly developed method are examined via simulation studies and the practical utility of our proposal is illustrated with the analysis of a lung cancer dataset.

Latent Variable Augmented Sparse Regression

Zemin Zheng

University of Science and Technology of China, China. *zhengzm@ustc.edu.cn*

Abstract: As a powerful tool for producing interpretable models, sparse modeling has gained popularity for analyzing large-scale data sets. Most of existing methods assume implicitly that all features in a model are observable. Yet some latent confounding factors may potentially exist in the hidden structure of the original model. On the other hand, the key assumption of sparsity that enables high-dimensional inference has been questioned in many applications. In this paper, we propose a new framework, latent variable augmented sparse regression (LAVAR), based on the conditional sparsity assumption that the coefficient vector is sparse after taking out the latent features. In particular, we consider one potential family of latent variables that are linearly dependent on a group of observable features, represented by the population principal components of the random matrix consisting of these features. The latent factors are estimated by the sample score vectors and asymptotic properties are established for a wide class of distributions. With the aid of these properties, we prove that high-dimensional thresholded regression with estimated latent variables can enjoy model selection consistency and oracle inequalities under various prediction and variable selection losses for both observable covariates and latent confounding factors. Our new method and results are evidenced by simulation and real data examples.

Wed, June 29 (08:30-10:10) l TCP32

Model Selection and Hypothesis Testing

Organizer: Jingheng Cai (Sun Yat-sen University)

Chair: Jingheng Cai (Sun Yat-sen University)

Wild Bootstrap Tests for Serial Correlation of Time Series Objects

Taewook Lee

Hankuk University of Foreign Studies, South Korea. *twlee@hufs.ac.kr*

Abstract: In this era of Big Data, large-scale data storage provides the motivation for statisticians to analyze new types of data. The proposed work concerns testing serial correlation in a sequence of sets of time series, here referred to as time series objects. An example is serial correlation of monthly stock returns when daily stock returns are observed. One could consider a representative or summarized value of each object to measure the serial correlation, but this approach would ignore information about the variation in the observed data. We develop Kolmogorov-Smirnov type tests with the standard bootstrap and wild bootstrap Ljung-Box test statistics for serial correlation in mean and variance of time series objects, which take the variation within a time series object into account. We study the asymptotic property of the proposed tests and present their finite sample performance using simulated and real examples.

Comparison of Non-nested Models Under Composite Kullback-Leibler Divergence

Chi Tim Ng

Chonnam National University, South Korea. *easterlyng@gmail.com*

Abstract: In this talk, it is illustrated that under the model misspecification cases, the closeness of two or more competing non-nested models to the truth can be compared under a procedure that is more general than that proposed in Vuong (1989). Composite Kullback-Leibler divergence can be used in such a procedure when the exact likelihood is difficult to compute. Large deviation theory is further used to obtain a bound of the power of rejecting the null hypothesis that the two models are equally close to the true model. Such a bound can be expressed in terms of a constant gamma belonging to [0,1) that can be computed empirically without any knowledge of the data generating mechanism. The constant gamma can be used to compare the ability of test procedures based on different composite likelihood to conclude a difference between two models.

A Criterion-based Model Comparison Statistic for Structural Equation models with Heterogeneous Data

Junhao Pan

Sun Yat-sen University, China. *panjunh@mail.sysu.edu.cn*

Abstract: Heterogeneous data are common in social, educational, medical and behavioral sciences. Recently, finite mixture structural equation models (SEMs) and two-level SEMs have been respectively proposed to analyze different kinds of heterogeneous data. Due to the complexity of these two kinds of SEMs, model comparison is difficult. For instance, the computational burden in evaluating the Bayes factor is heavy, and the Deviance Information Criterion may not be appropriate for mixture SEMs. In this paper, a Bayesian criterion-based method called the Lv measure, which involves a component related to the variability of the prediction and a component related to the discrepancy between the data and the prediction, is proposed. Moreover, the calibration distribution is introduced for formal comparison of competing models. Two simulation studies, and two applications based on real data sets are presented to illustrate the satisfactory performance of the Lv measure in model comparison.

Testing for the Equality of Integration Orders of Multiple Series

Man Wang

Donghua University, China. *wangman@dhu.edu.cn*

Abstract: Testing for the equality of integration orders is an important topic in time series analysis because it constitutes an essential step in testing for cointegration. However, for the case of multivariate fractionally integrated series, most tests are constrained to the stationary and invertible series, and some become invalid under the presence of cointegration or involve user-chosen parameters. Hualde (2013) overcomes those draw-backs with a residual-based test, which is restricted to the bi-variate case only. For the multiple series case, one reasonable extension of this test is a composition of testing for an array of bi-variate series, which is computationally demanding. In this paper, a simple one-step residual-based test is proposed for the multivariate series, which overcomes the aforementioned drawbacks and is easy to implement. Under certain regularity conditions, the test statistic has an asymptotic standard normal distribution under the null hypothesis of equal integration orders and diverges to infinity under the alternative. As reported in a Monte Carlo experiment, the proposed test possesses satisfactory sizes and powers.

Wed, June 29 (08:30-10:10) I CP06

Bayesian Session 1

Chair: Yan Liu (Ocean University of China)

Bayesian Predictive Distributions in Nonparametric Function Prediction

Keisuke Yano

The University of Tokyo, Japan. *keinet.marmar@gmail.com*

Abstract: Nonparametric function prediction is to predict a future function based on the past observation. In this talk, we provide constructions of the Bayesian predictive distributions that are asymptotically minimax under the Kullback—Leibler loss when the parameter space is an ellipsoid. We also provide constructions of the Bayesian predictive distributions that are adaptive asymptotically minimax under the Kullback—Leibler loss.

Approximate Bayesian Inference with Pseudo-likelihood

Ray S W Chung

The Hong Kong University of Science and Technology, Hong Kong. *swchungaa@ust.hk*

Abstract: Bayesian inference can effectively deal with a wide range of complicated statistical problems like high-dimensional inference, latent variable filtering and statistical learning. In classical Bayesian analysis, we need to fully specify the likelihood of underlying models so as to carry out statistical computation for posterior inference. The requirement of likelihood limits the application of Bayesian approach in solving semi-parametric problems or problems whose full likelihood is computationally intractable. In this paper, we propose an approximate Bayesian inference framework which incorporates pseudo-likelihood. It is expected that without the need of full likelihood specification, we can extend the scope of problems which Bayesian inference can solve. Two examples, the GARCH model and the spatiotemporal model, are taken to demonstrate our framework. Results in the examples show that the approximate Bayesian inference framework can provide both consistent estimates as well as good credible interval coverage.

Quantile Regression-based Bayesian Nonlinear Mixed-effects Joint Models for Survival-longitudinal Data with Multiple Features

Yangxin Huang

University of South Florida, United States. *yhuang@health.usf.edu*

Abstract: This paper explores Bayesian nonlinear mixed-effects joint models for quantiles of longitudinal response, mismeasured covariate and event time outcome with an attempt to (i) characterize the entire conditional distribution of the response variable based on quantile regression (QR) which may be more robust to outliers and mis-specification of error distribution, (ii) tailor accuracy from measurement error, evaluate non-ignorable missing observations and adjust departures from normality in covariate, and (iii) overcome shortages of confidence in specifying a time-to-event model. When statistical inference is carried out for a longitudinal data set with non-central location, non-linearity, non-normality, measurement error and missing values as well as event-time with being interval-censored, it is important to account for the simultaneous treatment of these data features in order to achieve more reliable and robust inferential results. Toward this end, we develop Bayesian joint modeling approach to simultaneously estimating all parameters in the three processes: QR-based nonlinear mixed-effects model for response using asymmetric Laplace distribution, linear mixed-effects model with skew-t distribution for mismeasured covariate in the presence of informative missingness and accelerated failure time model with unspecified nonparametric distribution for event time. We apply the proposed modeling approach to analyzing an AIDS clinical data set and conduct simulation studies to assess the performance of the proposed joint models and method.

Bayesian Spatial Temporal Model for Air Pollutant Data

Wai Ming Li

The Hong Kong University of Science and Technology, Hong Kong. *wmli@connect.ust.hk*

Abstract: Modeling and understanding air quality data has been an important topic for statistical applications in environmental science. Most researches focus on models either based on geographical locations or time dependence alone. In this paper, we introduce a new model that aims at modeling both spatial and temporal dependency using a latent variable structure. Specifically, time series dependency in latent variables is incorporated using a hierarchical Bayesian model with fixed effects of exogenous variables and random effects. We develop a full Bayesian estimation scheme for the model parameters. For illustration, a simulation study as well as a real data study based on air pollutant data of Hong Kong and neighboring region is given.

A Copula Based Multivariate Hierarchical Spatial Model with Applications to Daily Air Pollutant Extremes in Pearl River Delta

Ka Shing Chan

The Hong Kong University of Science and Technology, Hong Kong. *imkschan@ust.hk*

Abstract: Air pollution problem is a high priority concern in many cities. Poor air quality increases the risk of life-threatening conditions and burdens the public health care systems. A general approach to model extreme spatial events is through Bayesian hierarchical models. In the literature, this approach is mostly used to capture spatial dependence for only one type of event. This limits the applications to air pollutant data as different pollutants may chemically interact with each others. In this paper, we extend the Bayesian hierarchical models to a multivariate setting based on copulas so that our model is capable to handle both the spatial dependence and the dependence among multiple pollutants. We apply our proposed model to analyze daily maxima of nitrogen dioxide, ozone, respirable suspended particles (PM10) and fine suspended particles (PM2.5) collected in Pearl River Delta.

Wed, June 29 (08:30-10:10) | IP41

New Developments in High-Dimensional Spatial and Spatio-Temporal Modeling

Organizer: Chae Young Lim (Seoul National University)

Chair: Chae Young Lim (Seoul National University)

High Dimensional Variable Selection for Spatial Regression

Taps Maiti

Michigan State University, United States. *maiti@stt.msu.edu*

Abstract: Spatial regression is an important predictive tool in many scientific applications. When there are many available predictors, it is natural to select meaningful predictors. Penalized approach is proved to be effective for selecting a good model. It is noted that the variable selections for independent data do not work for spatially dependent data. Also the model selection research is still pre-mature in the context of high dimensional spatial regression. In this view, we consider the problem of selecting covariates in spatial regression model when it is possibly high-dimensional and/or the data could be non-Gaussian. In particular, we study various penalized approaches to deal with continuous spatial data and/or discrete spatial data. Several theoretical results are derived and their finite sample properties are investigated numerically.

Computational Challenges with Big Environmental Data

Marc G. Genton

King Abdullah University of Science and Technology, Saudi Arabia. *marc.genton@kaust.edu.sa*

Abstract: In this talk, we discuss two types of computational challenges arising from big environmental data. The first type occurs with multivariate or spatial extremes. Indeed, inference for max-stable processes observed at a large collection of locations is among the most challenging problems in computational statistics, and current approaches typically rely on less expensive composite likelihoods constructed from small subsets of data. We explore the limits of modern state-of-the-art computational facilities to perform full likelihood inference and to efficiently evaluate high-order composite likelihoods. With extensive simulations, we assess the loss of information of composite likelihood estimators with respect to a full likelihood approach for some widely-used multivariate or spatial extreme models, we discuss how to choose composite likelihood truncation to improve the efficiency, and we also provide recommendations for practitioners. The second type of challenges occurs with the emulation of climate model outputs. We consider fitting a statistical model to 1 billion global 3D spatio-temporal temperature data using a distributed computing approach. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a non-trivial model to a data set with a covariance matrix comprising of 10^{18} entries. The talk is based on joint work with Stefano Castruccio and Raphael Huser.

Hierarchical Low Rank Approximation for Large Spatial Datasets

Ying Sun

King Abdullah University of Science and Technology, Saudi Arabia. *ying.sun@kaust.edu.sa*

Abstract: Datasets in the fields of climate and environment are often very large and irregularly spaced. To model such datasets, the widely used Gaussian process models in spatial statistics face tremendous challenges due to the prohibitive computational burden. Various approximation methods have been introduced to reduce the computational cost. However, most of them rely on unrealistic assumptions of the underlying process and retaining statistical efficiency remains an issue. In this work, we develop a new approximation scheme for maximum likelihood estimation. We show how the composite likelihood method can be adapted to provide different types of hierarchical low rank approximations that are both computationally and statistically efficient. The improvement of the proposed method is explored theoretically; the performance is investigated by numerical and simulation studies; and practicality is illustrated through applying our methods to 2 million measurements of soil moisture in the area of Mississippi River basin, which facilitates better understanding of the climate variability.

Spatial Methods for Nonstationary Fields Using Compact Basis Functions

Soutir Bandyopadhyay

Lehigh University, United States. *sob210@lehigh.edu*

Abstract: Kriging is a non-parametric regression method used in geostatistics for estimating curve and surfaces and it forms the core of most statistical methods for spatial data. In climate science these methods are extremely useful for estimating how climate varies over a geographic region when the observational data is sparse or the computer model runs are limited. A statistical challenge is to implement spatial methods for large sample sizes and also account for the heterogeneity in the physical fields, both common features of many geophysical problems. Equally important is to provide companion measures of uncertainty so that the estimated surfaces can be compared in an objective way and are suitable for decision making. Here we present a statistical method that expands the spatial field in a large number of basis functions of varying resolution but all with compact support. Parsimonious models for the precision matrix of the basis coefficients are able to approximate standard covariance models but also scale to large numbers of spatial locations.

Wed, June 29 (10:30-12:10) l DL08 l Sponsor: Bernoulli Society

Heterogeneity in Large-Scale Data, with Connections to Causal Inference

Chair: Regina Y. Liu (Rutgers University)

Distinguished Lecturer: Peter Lukas Buhlmann (ETH, Zurich, Seminar for Statistics)

The Power of Heterogeneous Large-scale Data for High-dimensional Causal Inference

Peter Lukas Buhlmann

ETH, Zurich, Seminar for Statistics, Switzerland. *peter.buehlmann@stat.math.ethz.ch*

Abstract: We present a novel methodology for causal inference based on an invariance principle. It exploits the advantage of heterogeneity in larger datasets, arising from different experimental conditions (i.e. an aspect of "Big Data"). Despite fundamental identifiability issues, the method comes with statistical confidence statements leading to more reliable results than alternative procedures based on graphical modeling. We also discuss applications in biology, in particular for large-scale gene knock-down experiments in yeast where computational and statistical methods have an interesting potential for prediction and prioritization of new experimental interventions.

Penalized Estimation Methods for High-dimensional Causal Discovery

Ali Shojaie

University of Washington, United States. *ashojaie@uw.edu*

Abstract: Directed networks are widely used to model the interactions among components of complex social and biological systems, particularly when such interactions have causal interpretations. However, estimation of directed networks from observational data is in general challenging and the direction of the edges may not be estimable. On the other hand, interventional and time course data offer additional insight into causal interactions among high-dimensional variables. We will discuss applications of penalized estimation methods for learning directed networks. We will discuss asymptotic properties of the proposed estimators, and demonstrate their application using simulated and real-data examples.

Goodness of Fit Tests for High-dimensional Linear Models

Rajen Dinesh Shah

University of Cambridge, United Kingdom. *rds37@cam.ac.uk*

Abstract: In this talk, I will introduce a framework for constructing goodness of fit tests in both low and high-dimensional linear models. Our approach involves applying regression methods to the scaled residuals following either an ordinary least squares or Lasso fit to the data, and using some proxy for prediction error as the final test statistic. We call this family Residual Prediction (RP) tests. We show that simulation can be used to obtain the critical values for such tests in the low-dimensional setting, and demonstrate that some form of the parametric bootstrap can do the same when the high-dimensional linear model is under consideration. We show that RP tests can be used to test for significance of groups or individual variables as special cases, and here they compare favourably with state of the art methods, but we also argue that they can be designed to detect a variety of model misspecifications including heterogeneity among observations, heteroscedasticity and different forms of non-linearity.

Wed, June 29 (10:30-12:10) | DL16 | Sponsor: Australia and New Zealand

Variational Inference

Chair: Alan Welsh (Australian National University)

Distinguished Lecturer: Matt Wand (University of Technology Sydney)

Fast Approximate Inference for Arbitrarily Large Statistical Models via Message Passing

Matt Wand

University of Technology Sydney, Australia. *matt.wand@uts.edu.au*

Abstract: We explain how the notion of message passing can be used to streamline the algebra and computer coding for fast approximate inference in large Bayesian statistical models.

In particular, this approach is amenable to handling arbitrarily large models of particular types once a set of primitive operations is established. The approach is founded upon a message passing formulation of mean field variational Bayes that utilizes factor graph representations of statistical models. The notion of factor graph fragments is introduced and is shown to facilitate compartmentalization of the required algebra and coding.

Variational Approximations for Directional Data of Arbitrary Dimension

Jay Breidt

Colorado State University, United States. *jbreidt@stat.colostate.edu*

Abstract: In information retrieval and other contexts, directional data of arbitrary dimension arise as feature vectors are normalized to unit length and compared with one another. For example, a text document can be represented by a vector of word frequencies, and similarity between two documents can be measured by the cosine of the angle between their respective vectors. Equivalently, the two vectors can be regarded as points on the unit hypersphere. The general projected normal distribution is a simple and intuitive model for directional data in any dimension: a multivariate normal random vector divided by its length is the projection of that vector onto the surface of the unit hypersphere. Observed data consist of the projections, but not the lengths. Until recently, inference for this model has been restricted to the two-dimensional (circular) case, using Bayesian methods with data augmentation to generate the latent lengths and a Metropolis-within-Gibbs algorithm to sample from the posterior. A new parameterization of the general projected normal distribution makes inference in any dimension tractable. Under this new parameterization, the full conditionals of the unknown parameters have closed forms, and Gibbs sampling is fast and easy. However, convergence of the sampler slows with increasing dimension of the directional data or increasing sample size. Variational approximations are attractive in this context, but the standard mean-field approach breaks down if it uses the latent lengths and assumes their independence in the posterior. We explore alternative variational approximations that use a closed-form expression for the likelihood of the general projected normal. We compare results of the Gibbs sampler and the variational approximation in some 10-dimensional and 50-dimensional cases.

This is joint work with Daniel Hernandez-Stumpfhauser, University of North Carolina-Chapel Hill.

Flexible Online Multivariate Regression with Variational Bayes and the Matrix-variate Dirichlet Process

Meng Hwee Victor Ong

National University of Singapore, Singapore. *victor84@u.nus.edu.sg*

Abstract: In this presentation, we explore a flexible regression method where the distribution of a response vector changes with covariates. We use the matrix-variate Dirichlet process as a prior for a mixing distribution on a coefficient in a multivariate linear regression model. The method is attractive as it allows for borrowing strength across different component regressions. We develop fast online approaches to fitting the model using variational Bayes methods for application to large datasets. We also propose a regression adjustment approach to improving the predictive performance of the online algorithm.

Wed, June 29 (10:30-12:10) I IP11 I Sponsor: ICSA

Nonparametric Modelling and High Dimensional Data Analysis

Organizer: Wenyang Zhang (University of York)

Chair: Zhezhen Jin (Columbia University)

Empirical Bayes Prediction for the Multivariate Newsvendor Loss Function

Gourab Mukherjee

University of Southern California, United States. *gmukherj@marshall.usc.edu*

Abstract: Motivated by an application in inventory management, we consider the multi-product newsvendor problem of finding the optimal stocking levels that minimize the total backorder and lost sales costs. We focus on a setting where we have a large number of products and observe only noisy estimates of the underlying demand. We develop an Empirical Bayes methodology for predicting stocking levels using data-adaptive linear shrinkage strategies, which are constructed by minimizing uniformly efficient asymptotic risk estimates. In calculating the magnitude and direction of shrinkage, our proposed predictive rules incorporate the asymmetric nature of the piecewise linear newsvendor loss function and are shown to be asymptotically optimal. Using simulated data, we study the non-asymptotic performance of our method and obtain encouraging results.

Low-dimensional Confounder Adjustment and High-dimensional Penalized Estimation for Survival Analysis

Jialiang Li

National University of Singapore, Singapore. *stalj@nus.edu.sg*

Abstract: High-throughput profiling is now common in biomedical research. In this paper we consider the layout of an etiology study composed of a failure time response, and gene expression measurements. In current practice, a widely adopted approach is to select genes according to a preliminary marginal screening and a follow-up penalized regression for model building. Confounders, including for example clinical risk factors and environmental exposures, usually exist and need to be properly accounted for. We propose covariate-adjusted screening and variable selection procedures under the accelerated failure time model. While penalizing the high-dimensional coefficients to achieve parsimonious model forms, our procedure also properly adjust the low dimensional confounder effects to achieve more accurate estimation of regression coefficients. We establish the asymptotic properties of our proposed methods and carry out simulation studies to assess the finite sample performance. Our methods are illustrated with a real gene expression data analysis where proper adjustment of confounders produces more meaningful results.

Innovated Interaction Screening for High-dimensional Nonlinear Classification

Daoji Li

University of Central Florida, United States. *Daoji.Li@ucf.edu*

Abstract: This paper 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. This is a joint work with Yingying Fan, Yinfei Kong and Zemin Zheng.

High-dimensional A-learning for Optimal Dynamic Treatment Regimes

Rui Song

North Carolina State University, United States. *rsong@ncsu.edu*

Abstract: Precision medicine is a medical paradigm that focuses on finding the most effective treatment decision based on individual patient information. For many complex diseases, such as cancer, treatment decisions need to be tailored over time according to patients' responses to previous treatments. Such an adaptive strategy is referred as a dynamic treatment regime. A major challenge in deriving an optimal dynamic treatment regime arises when an extraordinary large number of prognostic factors, such as patient's genetic information, demographic characteristics, medical history and clinical measurements over time are available, but not all of them are necessary for making treatment decision. This makes variable selection an emerging need in precision medicine.

We propose a penalized multi-stage A-learning for deriving the optimal dynamic treatment regime when the number of covariates is of the non-polynomial (NP) order of the sample size. To preserve the double robustness property of the A-learning method, we adopt the Dantzig selector which directly penalizes the A-leaning estimating equations. Oracle inequalities of the proposed estimators for the parameters in the optimal dynamic treatment regime and error bounds on the difference between the value functions of the estimated optimal dynamic treatment regime and the true optimal dynamic treatment regime are established. Empirical performance of the proposed approach is evaluated by simulations and illustrated with an application to a data from the STAR*D study.

Wed, June 29 (10:30-12:10) l IP22 l Sponsor: China

Recent Advances in Complex Data Analysis

Organizer: Liping Zhu (Renmin University of China)

Chair: Zhongyi Zhu (Fudan University)

Empirical Likelihood Inference in Linear Regression with Nonignorable Missing Response

Wangli Xu

Renmin University of China, China. *xwlbnu@163.com*

Abstract: Parameter estimation for nonignorable nonresponse data is a challenging issue as the missing mechanism is unverified in practice and the parameters of response probabilities need to be estimated. This article aims at applying the empirical likelihood to construct the confidence intervals for the parameters of interest in linear regression models with nonignorable missing response data and the nonignorable missing mechanism is specified as an exponential tilting model. Three empirical likelihood ratio functions based on weighted empirical likelihood and imputed empirical likelihood are defined. It is proved that, except one that is chi-squared distributed, all the others are asymptotically weighted chi-squared distributed whenever the tilting parameter is either given or estimated. The asymptotic normality for the related parameter estimates is also investigated. Simulation studies are conducted to evaluate the finite sample performance of the proposed estimates in terms of coverage probabilities and average widths for the confidence intervals of parameters. A real data analysis is analyzed for illustration.

On Marginal Sliced Inverse Regression for Ultrahigh Dimensional Model-free Feature Selection

Zhou Yu

East China Normal University, China. *zyu@stat.ecnu.edu.cn*

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 ultrahigh 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.

Ensemble Sufficient Dimension Folding Methods on Analyzing Matrix-valued Data

Yuan Xue

University of International Business and Economics, China. *yuanxue@uibe.edu.cn*

Abstract: In this paper, we construct novel sufficient dimension folding methods on analyzing matrix-valued objects. Different from conventional vector-valued predictor, the predictor is a random matrix. Traditional dimension reduction methods fail to preserve the matrix structure of the reduced predictor. Dimension folding methods for matrix-/array-valued predictor 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 structural dimensions of CDFS. Simulated examples demonstrate the performances of the folded ensemble methods by comparing with existing inverse dimension folding methods. The efficacy of proposed folded-MAVE ensemble method is evaluated by comparing with the inverse methods on analyzing a matrix-valued data.

Efficient and Nonparametric Causal Inference

Zheng Zhang

Renmin University of China, China. *zzhang1989@gmail.com*

Abstract: The estimation of average treatment effects based on observational data is extremely important in practice and has been studied by generations of statisticians under different frameworks. Existing globally efficient estimators require non-parametric estimation of a propensity score function, an outcome regression function or both, but their performance can be poor in practical sample sizes. Without explicitly estimating either function, we consider a wide class of calibration weights constructed to attain an exact three-way balance of the moments of observed covariates among the treated, the control and the combined group. The wide class includes exponential tilting, empirical likelihood and generalized regression as important special cases, and extends survey calibration estimators to different statistical problems and with important distinctions. Global semiparametric efficiency for the estimation of average treatment effects is established for this general class of calibration estimators. The results show that efficiency can be achieved by solely balancing the covariate distributions without resorting to direct estimation of the propensity score or outcome regression function. We also propose a consistent estimator for the efficient asymptotic variance, which does not involve additional functional estimation of either the propensity score or the outcome regression functions. The variance estimator proposed outperforms existing estimators that require a direct approximation of the efficient in influence function. This is a joint work with Gary Chan and Phillip Yam.

Wed, June 29 (10:30-12:10) I IP43

Statistical Methodology for Biomedical Sciences

Organizer: Xuming He (University of Michigan)

Chair: Mi-Ok Kim (Cincinnati Children's Hospital Medical Center)

Approximate Median Regression for Complex Survey Data with Skewed Response

Stuart Lipsitz

Harvard University, United States. *slipsitz@partners.org*

Abstract: The ready availability of public-use data from various large national complex surveys has immense potential for the assessment of frequency, prevalence, treatment options and their effectiveness. In addition, complex surveys can be used to identify risk factors for important diseases such as cancer. Existing statistical methods based on estimating equations and/or utilizing resampling methods are often not valid with survey data due to complex design features, i.e. stratification, multistage sampling and weighting. In this paper, we accommodate these design features in the analysis of highly skewed response variables arising from large complex surveys. Specifically, we propose a double-transform-both-sides (DTBS) based estimating equations approach to estimate the median regression parameters of the highly skewed response; the DTBS approach applies the same Box-Cox type transformation twice to both the outcome and regression function. The usual sandwich variance estimate can be used in our approach, whereas a resampling approach would be needed for a pseudo-likelihood based on minimizing absolute deviations (MAD). Furthermore, the approach is relatively robust to the true underlying distribution, and has much smaller mean square error than a MAD approach. The method is motivated by an analysis of laboratory data on urinary iodine (UI) concentration from the National Health and Nutrition Examination Survey.

Identification of Homogeneous and Heterogeneous Variables in Pooled Cohort Studies

Mengling Liu

NYU School of Medicine, United States. *Mengling.Liu@nyumc.org*

Abstract: Pooled analyses utilize data from multiple studies and intend to achieve a large sample size for increased statistical power. When heterogeneity exists in variables' effects across studies, the simple pooling strategy fails to present a fair and complete picture of the effects of heterogeneous variables. Thus, it is important to investigate the homogeneous and heterogeneous structure of variables in pooled studies. In this talk, I will present our recent work on using composite penalty regularized likelihood approaches to identifying variables with heterogeneous effects in pooled studies. The methods will be demonstrated using numerical simulations and real study applications.

Conditional Graphical Models with Applications in Integrative Genomics

Jie Peng

University of California, Davis, United States. *jiepeng@ucdavis.edu*

Abstract: Motivated by network construction in integrative genomics, we developed a sparse conditional graphical model (CG model) called Spacemap. Spacemap integrates two high dimensional data types through a penalized multivariate regression framework in the high-dimension-low-sample-size regime. It learns an undirected graph among one type of nodes (referred to as the responses) together with a directional graph encoding how another type of nodes (referred to as the covariates) perturb the responses. For example, the responses could be mRNA expressions or protein expressions and the covariates could be gene copy numbers. Strategies for model fitting and model selection will be discussed. Simulation studies show that Spacemap has superior power over graphical models where the two types of nodes are not differentiated and a likelihood-based CG model. We will further demonstrate Spacemap on a cancer data set with mRNA expressions, protein expressions and copy number variations.

Test for Genomic Imprinting Effects on the X Chromosome

Wing Kam Fung

The University of Hong Kong, Hong Kong. *wingfung@hku.hk*

Abstract: Methods for detecting imprinting effects have been developed primarily for autosomal markers. However, no method is available in the literature to test for imprinting effects on the X chromosome. Therefore, it is necessary to suggest methods for detecting such imprinting effects. In this talk, the parental-asymmetry test on the X chromosome (XPAT) is first developed to test for imprinting for qualitative traits in the presence of association, based on family trios each with both parents and their affected daughter. Then, we propose 1-XPAT to tackle parent-daughter pairs, each with one parent and his/her affected daughter. By simultaneously considering family trios and parent-daughter pairs, C-XPAT is constructed to test for imprinting. Further, we extend the proposed methods to accommodate complete (with both parents) and incomplete (with one parent) nuclear families having multiple daughters of which at least one is affected. Simulations are conducted to assess the performance of the proposed methods under different settings. Simulation results demonstrate that the proposed methods control the size well, irrespective of the inbreeding coefficient in females being zero or nonzero. By incorporating incomplete nuclear families, C-XPAT is more powerful than XPAT using only complete nuclear families. For practical use, these proposed methods are applied to analyze the rheumatoid arthritis data.

Wed, June 29 (10:30-12:10) I IP52

Challenges and Recent Advances in Methods for Missing Data Problems

Organizer: Zonghui Hu (Biostatistics Research Branch, DCR)

Chair: Chiung-Yu Huang (Johns Hopkins University)

Improved Estimation of Average Treatment Effects on the Treated: Local Efficiency, Double Robustness, and Beyond

Zhiqiang Tan

Rutgers University, United States. *ztan@stat.rutgers.edu*

Abstract: Estimation of average treatment effects (ATT) on the treated is an important topic in causal inference. But this problem seems to be often treated as a simple modification or extension of the problem of estimating overall average treatment effects (ATE). In this talk, we will examine semiparametric theory for estimation of ATT, with interesting comparison with estimation of ATE, and use such theory to derive augmented inverse probability weighted (AIPW) estimators that are locally efficient and doubly robust. Moreover, we develop calibrated regression and likelihood estimators that are not only locally efficient and doubly robust, but also intrinsically efficient in achieving greater efficiency than AIPW estimators when a propensity score model is correctly specified but an outcome regression model may be misspecified. Finally, we present simulation studies and a real application to demonstrate the advantage of the proposed methods when compared with existing methods.

Bayesian Pattern Mixture Models for the Analysis of Repeated Attempt Designs

Michael Daniels

The University of Texas at Austin, United States. *mjdaniels@austin.utexas.edu*

Abstract: It is not uncommon in follow-up studies to make multiple attempts to collect a measurement after baseline. Recording whether these attempts are successful or not provides useful information for the purposes of assessing the missing at random (MAR) assumption and facilitating missing not at random (MNAR) modeling. This is because measurements from subjects who provide this data after multiple failed attempts may differ from those who provide the measurement after fewer attempts. This type of 'continuum of resistance' to providing a measurement has hitherto been modeled in a selection model framework, where the outcome data is modeled jointly with the success or failure of the attempts given these outcomes. We present a Bayesian pattern mixture approach to model this type of data. We re-analyze the repeated attempt data from a trial that was previously analyzed using a selection model approach. Our pattern mixture model is more flexible and transparent than the models that have previously been used to model repeated attempt data and allows for sensitivity analysis and informative priors.

Semiparametric Pseudoscore Estimation for Regressions with Potentially High-dimensional But Incompletely Observed Covariates

Zonghui Hu

National Institutes of Health, United States. *zonghui.hu@nih.gov*

Abstract: We study the parametric regression with a vector of fully observed regressors *Z* and an incompletely observed regressor *X*. To handle missing covariate data, maximum likelihood estimation (MLE) via expectation-maximisation (EM) is the most efficient. However, the MLE is sensitive to the assumed conditional distribution of *X*: misspecification leads to inconsistent estimation of the regression parameters. Under missing at random assumption, we propose an EM type estimation via a semiparametric pseudoscore, where the pseudoscore is estimated nonparametrically through a parametric working index. The working index targets the conditional mean score given *Z* and the response and is based upon some preliminary distributional information about *X*. Nonparametric regression estimation of the pseudoscore entails robustness. Meanwhile, it is free of the curse of dimensionality even when *Z* is high-dimensional, due to the adoption of the parametric working index which is one-dimensional for each parameter. The proposed estimator is more than doubly robust: it is consistent if either the pattern of missingness in *X* is correctly specified or the working index is appropriately specified, and it attains the optimal efficiency when both conditions are satisfied. Since specification of the conditional distribution of *X* is challenging in most applications, the proposed is more practical than the conventional EM. Numerical performances are explored by simulations and an example in hepatitis C study.

Nonparametric Sufficient Dimension Reduction with Missing Predictors at Random

Qihua Wang

Chinese Academy of Sciences, China. *qhwang@amss.ac.cn*

Abstract: In some practical problems, a subset of predictors is subject to missingness when the dimensionality of the predictor vector is high. In this situation the standard sufficient dimension reduction (SDR) methods cannot be applied to avoiding the "curse of dimensionality" problem. In this paper, a nonparametric procedure is developed to handle the dimension-reduction problems with predictors missing at random. The sliced inverse regression (SIR) is used to illustrate this procedure. It is shown that the proposed estimator of dimension reduction directions is asymptotically normal under some mild conditions. And the finite-sample performances of the proposed method are evaluated through comprehensive simulations and a real data analysis. Furthermore, the general idea of the proposed nonparametric procedure is extend to other two popular SDR methods, namely sliced average variance (SAVE) and principal Hessian direction (PHD).

Wed, June 29 (10:30-12:10) l TCP11

New Statistical Methods for Genetic Data Analysis

Organizer: Anne Buu (University of Michigan)

Chair: Runze Li (The Pennsylvania State University)

Robust Modeling of RNA-Seq Data

Hui Jiang

University of Michigan, United States. *jianghui@umich.edu*

Abstract: In recent years, datasets from high-throughput sequencing of transcriptomes (RNA-Seq) with hundreds of individuals have been generated from many large projects. Combining these gene expression data with genotype data, one can learn how gene expression levels are associated with genetic variations. However, the technical nature of RNA-Seq makes it very difficult to obtain accurate measurements of gene expression levels. In this talk, we will introduce some recently develop methods for the robust modeling of RNA-Seq data, which facilitates more accurate downstream analyses such as the detection of eQTLs and differentially expressed genes.

Random Field Modelling of Genetic Association for Sequencing Data in Family-based Studies

Ming Li

Indiana University, United States. *li498@indiana.edu*

Abstract: Emerging studies using next-generation sequencing technology hold great promise for the identification and fine mapping of novel genetic variants, especially rare variants, contributing to complex human diseases. However, detecting these disease-susceptibility rare variants remains a great challenge because of the heterogeneous nature and low frequency of rare variants. Multiple rare variants within the same gene can independently influence the disease (i.e., allelic heterogeneity), and rare variants in different genes can also be involved in related pathways underlying diseases (i.e., locus heterogeneity). Advanced analytical methods are in great need to account for the genetic heterogeneity of complex human diseases. In this talk, we will introduce a family-based genetic random filed (FGRF) method for association analyses of sequencing data in family-based association studies. By utilizing information from family members, the proposed method is robust to population stratification and gains improved performance in presence of genetic heterogeneity. The proposed method is compared to other existing methods through simulation studies and real data applications for investigating the genetic etiology of complex diseases/traits.

The Genetic Architecture of Complex Phenotypes: New Insight from Game Theory

Rongling Wu

The Pennsylvania State University, United States. *rwu@phs.psu.edu*

Abstract: Despite their paramount importance to life sciences, our understanding of quantitatively inherited traits is very limited because they involve so many complex genetic and physiological mechanisms. In this talk, I present a new theory for mapping complex traits by integrating game theory into the statistical framework of genetic association studies. In nature, the development of any trait is never an isolated process, rather than it encompasses a web of interactions between its internal components and external interfaces through Darwinian natural selection. This universal principle, quantified by game theory, is utilized and incorporated to identify causal intermediate pathways between genotype and phenotype. The game-based mapping strategy established by a group of differential equations breaks through traditional ways to dissect trait phenotypes without taking account into the underlying ecological and evolutionary mechanisms. The new strategy can not only map quantitative traits more precisely and more efficiently, but also provide an unprecedented tool to study the genetic control of biological processes in evolutionary, ecological and biomedical studies.

Estimation of stratified Mark-specific Proportional Hazards Models under Two-phase Sampling with Application to HIV Vaccine Efficacy Trials

Guangren Yang

Jinan University, China. *tygr@jnu.edu.cn*

Abstract: This article develops estimation procedures for the stratified mark-specific proportional hazards model under two-phase sampling where the baseline functions may vary with strata. The mark-specific proportional hazards model has been studied to evaluate mark-specific relative risks where the mark is the genetic distance of an infecting HIV sequence to an HIV sequence represented inside the vaccine. This research is motivated by preventive HIV vaccine efficacy trials, for assessing the association of vaccine-induced immune response biomarkers, which are measured via two-phase sampling for efficiency sake, on the mark-specific incidence of acquiring HIV infection. The augmented inverse probability weighted complete-case estimation methods are developed. The asymptotic properties of the proposed estimators are derived, and their finite-sample performances are examined in a comprehensive simulation study. The methods are shown to have satisfactory performance, and are applied to the RV144 vaccine trial to assess whether immune response correlates of HIV infection are stronger for HIV infecting sequences similar to the vaccine than for sequences distant from the vaccine.

An Efficient Genome-wide Association Test for Multivariate Phenotypes Based on the Fisher Combination Function

James Yang

University of Michigan, United States. *jjyang@umich.edu*

Abstract: In genome-wide association studies (GWAS) for complex diseases, the association between a SNP and each phenotype is usually weak. Combining multiple related phenotypic traits can increase the power of gene search and thus is a practically important area that requires methodology work. This study provides a comprehensive review of existing methods for conducting GWAS on complex diseases with multiple phenotypes including the multivariate analysis of variance (MANOVA), the principal component analysis (PCA), the generalizing estimating equations (GEE), the trait-based association test involving the extended Simes procedure (TATES), and the classical Fisher combination test. We propose a new method that relaxes the unrealistic independence assumption of the classical Fisher combination test and is computationally efficient. To demonstrate applications of the proposed method, we also present the results of statistical analysis on the Study of Addiction: Genetics and Environment (SAGE) data.

Wed, June 29 (10:30-12:10) | TCP26

Statistical Issues in Analyzing Bioinformatics Data

Organizer: Xiaodan Fan (The Chinese University of Hong Kong)

Chair: Han Li (Shenzhen University)

Lengthening and Shortening of Tumour-derived Plasma DNA: Reconciling a Long-standing Controversy

Peiyong Jiang

The Chinese University of Hong Kong, Hong Kong. *jiangpeiyong@cuhk.edu.hk*

Abstract: The analysis of tumor-derived circulating cell-free DNA opens up new possibilities for performing liquid biopsies for the assessment of solid tumors. Although its clinical potential has been increasingly recognized, many aspects of the biological characteristics of tumor-derived cell-free DNA remain unclear. With respect to the size profile of such plasma DNA molecules, a number of studies reported the finding of increased integrity of tumor-derived plasma DNA, whereas others found evidence to suggest that plasma DNA molecules released by tumors might be shorter. Here, we performed a detailed analysis of the size profiles of plasma DNA in 90 patients with hepatocellular carcinoma, 67 with chronic hepatitis B, 36 with hepatitis B-associated cirrhosis, and 32 healthy controls. We used massively parallel sequencing to achieve plasma DNA size measurement at single-base resolution and in a genome-wide manner. Tumor-derived plasma DNA molecules were further identified with the use of chromosome arm-level z-score analysis (CAZA), which facilitated the studying of their specific size profiles. We showed that populations of aberrantly short and long DNA molecules existed in the plasma of patients with hepatocellular carcinoma. The short ones preferentially carried the tumor-associated copy number aberrations. We further showed that there were elevated amounts of plasma mitochondrial DNA in the plasma of hepatocellular carcinoma patients. Such molecules were much shorter than the nuclear DNA in plasma. These results have improved our understanding of the size profile of tumor-derived circulating cell-free DNA and might further enhance our ability to use plasma DNA as a molecular diagnostic tool.

Estimating Reproducibility in Genome-wide Association Studies

Weichuan Yu

The Hong Kong University of Science and Technology, Hong Kong. *eeyu@ust.hk*

Abstract: Genome-wide association studies (GWAS) are widely used to discover genetic variants associated with diseases. Replication study is a common verification method by using independent samples. An association is regarded as true positive with a high confidence when it can be identified in both primary study and replication study. Currently, there is no systematic study on the behavior of positives in the replication study when the positive results of primary study are considered as the prior information.

This talk proposes two probabilistic measures, the Reproducibility Rate (RR) and the False Irreproducibility Rate (FIR), to quantitatively describe the behavior of primary positive associations (i.e. positive associations identified in the primary study) in the replication study. RR is a conditional probability measuring how likely a primary positive association will also be positive in the replication study. This can be used to guide the design of replication study, and to check the consistency between the results of primary study and those of replication study. FIR, on the contrary, measures how likely a primary positive association may still be a true positive even when it is negative in the replication study. This can be used to generate a list of potentially true associations among the irreproducible findings for further scrutiny. Estimation methods for these two measures are given. Simulation results and real experiments show that our estimation methods have high accuracy and good prediction performance.

This is a joint work with Wei Jiang and Jinghao Xue.

DNA Methylation in Enhancers

Jiangwen Zhang

The University of Hong Kong, Hong Kong. *jzhang1@hku.hk*

Abstract: Cancer, a complex and fatal disease, is the result of combined genetic and epigenetic alteration. DNA methylation is the most common epigenetic mechanism in the mammalian genome, with cytosine methylation on CpG sites to regulate cognate gene expression. During tumorigenesis, DNA methylation patterns are widely altered compared with normal genomes. Most previous studies just focused on aberrant methylation of CpG islands around promoters. The potential function of expression-associated methylation away from gene promoters have not been fully studied. Alteration of DNA methylation in non-coding regions may happen on other regulatory sites, like enhancers. Enhancer demethylation can result in transcription factor binding, DNA looping and expression of eRNAs. These processes consequently may facilitate transcription of target genes.

In order to understand the methylation alteration on non-coding region in cancer, we propose a new method to infer the long-range epigenetic regulatory networks. Our method is based on information theory, the "mutual information". After applying this method to gene expression and DNA methylation in cancers from TCGA, we have discovered many regulatory links between transcriptome and methylome. More importantly, we can eliminate the indirect links with only direct regulations kept. And these findings can help prioritize the epigenetic events in down-regulation of tumor suppressors, and identify the pivotal factors for tumor initiation and progression. Some of our findings can be verified by some experiments in very recent studies. The preliminary results indicate a promising future for this novel method in cancer systems biology.

Tumor Purity and DMR Estimation from DNA Methylation Data

Xiaoqi Zheng

Shanghai Normal University, China. *xqzheng@shnu.edu.cn*

Abstract: DNA methylation is an important epigenetic mark controlling gene expression, thus playing pivotal roles in many cellular processes including embryonic development, genomic imprinting, X-chromosome inactivation, transposable element repression, and preservation of chromosome stability. Aberrant DNA methylations are known to be associated with human diseases such as cancers, lupus, muscular dystrophy, and imprinting related birth defects. Here, we proposed statistical models to infer tumor purity from DNA methylation Bisulfite sequencing and Inifinium 450K array data. We discover that in cancer samples, the distributions of data from Illumina Infinium 450k methylation microarray are highly correlated with tumor purities. We develop a simple but effective method to estimate purities from the microarray data. Analyses of the Cancer Genome Atlas data demonstrate favorable performance of the proposed method.

Wed, June 29 (10:30-12:10) l TCP29

Recent Advances on Random Processes and Related Problems

Organizer: Zhonggen Su (Zhejiang University)

Chair: Zhonggen Su (Zhejiang University)

The Fourth Moment Theorem for the Complex Multiple Wiener-Itô Integrals

Yong Liu

Peking University, China. *liuyong@math.pku.edu.cn*

Abstract: In this talk, we give a product formula of Hermite polynomials and show the relation between the real Wiener-Itô chaos and the complex Wiener-Itô chaos (or: multiple integrals), and then we extent the fourth moment theorem (or say: Nualart-Peccati criterion) to the complex multiple Wiener-Itô integrals. We also prove that the complex Hermite polynomials (or say: Hermite-Laguerre-Itô polynomials) are the eigenfunctions of complex Ornstein-Uhlenbeck operators. This talk is based on the following joint articles with Yong CHEN.

[1] Chen, Y., Liu Y., On the eigenfunctions of the complex Ornstein-Uhlenbeck operators, Kyoto J. Math., Vol. 54(3), 577-596, (2014)

[2] Chen Y., Liu Y., On the fourth Moment Theorem for the Complex Multiple Wiener-Itô Integrals, Preprint (2015).

Volume Growth and Escape Rate of Symmetric Diffusion Processes

Shun-Xiang Ouyang

South University of Science and Technology of China, China. *ouyangshx@hotmail.com*

Abstract: We give an upper rate function, in terms of the volume growth of the underlying state space, for the symmetric diffusion process associated with a symmetric, strongly local regular Dirichlet form. It extends the main result of Hsu and Qin [Ann. Probab. 38(4) 2010], where an upper rate function was given for Brownian motion on Riemannian manifold.

Double Contour Integral Formulas in Two Matrix Model and Related Non-intersecting Brownian Motions

Dong Wang

National University of Singapore, Singapore. *dongwangunc@gmail.com*

Abstract: The investigation of local statistics in two matrix model has seen substantial progress in last decade, thanks to the development of large size Riemann-Hilbert problems. The state of the art is the solution of the critical case of the model where the two matrices are with quadratic and symmetric quartic potentials each, and it is based on the 4x4 tacnode Riemann-Hilbert problem. In this talk, I will present an alternative approach to the two matrix model where one matrix is with quadratic potential and the other with arbitrary potential, based on double contour integral formulas involving the classical 2x2 Riemann-Hilbert problem that is associated to the one matrix model.

This talk is based on joint work with Tom Claeys, Arno Kuijlaars, and Karl Liechty.

The Stochastic Logarithmic Schrödinger Equation

Deng Zhang

Shanghai Jiao Tong University, China. *zhangdeng@amss.ac.cn*

Abstract: In this talk we will present the global existence and uniqueness of solutions to the stochastic logarithmic Schrödinger equation with linear multiplicative noise. The approach is mainly based on the rescaling approach and the method of maximal monotone operators. In addition, uniform estimates of solutions in the energy space H1 and in an appropriate Orlicz space are also obtained.

Metric Entropy of High Dimensional Convex Functions

Fuchang (Frank) Gao

University of Idaho, United States. *fuchang@uidaho.edu*

Abstract: Let K be a d-dimensional bounded closed convex set with non-empty interior, and let C(K) be the class of convex functions on K with L^r norm bounded by 1. We obtain sharp estimates of the entropy of C(K) under L^p metrics, $1<p<r$. In particular, the results imply that the universal lower bound $\varepsilon^{-d/2}$ is attained by all d-dimensional polytopes. While a general convex body can be approximated by inscribed polytopes, the entropy rate does not carry over to the limiting body. For example, if K is the closed unit ball, then the metric entropy has an higher order. The results have applications to questions concerning rates of convergence of nonparametric estimators of high-dimensional shape-constrained functions.

Wed, June 29 (10:30-12:10) | TCP30

Statistical Inference Under Model Uncertainties

Organizer: Stephen Lee (The University of Hong Kong)

Chair: Stephen Lee (The University of Hong Kong)

On a General Procedure for Constructing Confidence Sets under Partially Identified Models

Han Jiang

The University of Hong Kong, Hong Kong. *shirleyjiang@hku.hk*

Abstract: Recent years have seen a growing body of literature focusing on partially identified models, where observable data and credible assumptions can only identify the parameter of interest with a set, called the identified set, rather than a singleton. Manski (2003) provides a recipe of partial identification problems which may not be readily amenable to conventional statistical approaches. For certain partially identified models such as those defined using moment inequalities, new algorithms have been developed for constructing confidence sets for identified sets. The problem remains, however, unsolved outside such restrictive contexts. In the present study we propose a general confidence procedure which not only generalizes existing algorithms but also finds applications to settings as yet unexplored. The main thrust of our proposed procedure lies in an expansion step, by which we construct a family of nested sets and perform inferences centred on an expansion index. Wide choices of expansion indices and their corresponding nested sets allow of the flexibility and applicability necessary for dealing with a more general class of partially identified models. To illustrate the generality of our procedure, a simulation study is presented concerning the least quantile of squares in a partially identified model setting, a problem to which no solution has yet been found in the literature.

Variable Selection in Single-index Varying Coefficient Models

Anna Liu

University of Massachusetts Amherst, United States. *anna@math.umass.edu*

Abstract: Single index varying coefficient model is an attractive statistical model with its ability to hand high dimensional data and its ease of interpretation. Motivated by a Geoscience project and a TV rating project from the advertisement industry, we study the problem of index variable selection in the single index varying coefficient model. We consider both regression and classification problems, and use LASSO type of penalty for the variable selection purpose. We propose a new and easy-to-implement algorithm for the optimization which consists of two steps alternating between estimating the coefficient functions, and selecting/estimating the single index. We illustrate our algorithm with the above mentioned applications and our R package.

Quadratic Discriminant Analysis for High-dimensional Data

Yilei Wu

University of Waterloo, Canada. *y335wu@uwaterloo.ca*

Abstract: High-dimensional classification is an important and challenging statistical problem. We consider quadratic discriminant rules which simplify matrix structure instead of requiring sparsity assumptions --- either on the covariance matrices of each class (or their inverses), or on the standardized between-class distance. Under moderate conditions on the eigenvalues of population covariance matrices, our rules enjoy good asymptotic properties. Computationally, they are easy to implement and do not require large-scale mathematical programming. Numerically, they perform well in finite dimensions and with finite sample sizes. We also present real-data analyses of several classical micro-array data sets.

Spherical Cap Packing Asymptotics and Rank-extreme Detection

Kai Zhang

The University of North Carolina at Chapel Hill, United States. *zhangk@email.unc.edu*

Abstract: We study the spherical cap packing problem with a probabilistic approach. Such probabilistic considerations result in an asymptotic universal uniform sharp bound on the maximal inner product between any set of unit vectors and a stochastically independent uniformly distributed unit vector. When the set of unit vectors are themselves independently uniformly distributed, we further develop the extreme value distribution limit of the maximal inner product, which characterizes its stochastic uncertainty around the bound.

As applications of the above asymptotic results, we derive (1) an asymptotic universal uniform sharp bound on the maximal spurious correlation, as well as its uniform convergence in distribution when the explanatory variables are independently Gaussian; and (2) a sharp universal bound on the maximum norm of a low-rank elliptically distributed vector, as well as related limiting distributions. With these results, we develop a fast detection method for a low-rank in high dimensional Gaussian data without using the spectrum information.

Wed, June 29 (10:30-12:10) l CP08

Probability Session 1

Chair: Kouji Tahata (Tokyo University of Science)

A New Two-sample Test for High-dimension, Low-sample-size Data

Aki Ishii

University of Tsukuba, Japan. *ishii-akitk@math.tsukuba.ac.jp*

Abstract: A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. We call such data HDLSS data. In HDLSS data situations, one cannot claim any theoretical guarantee about accuracies in multivariate analysis. Hence, new theories and methodologies are required to develop for HDLSS data. One of the approaches is to study geometric representations of high-dimensional data. Yata and Aoshima (2012, JMVA) found two geometric representations of the eigenspace of an HDLSS sample covariance matrix and created the noise-reduction methodology which is a new PCA for high-dimensional data.

In this talk, we consider two-sample tests in HDLSS situations. We pay special attention to the first principal component when the dimension goes to infinity while the sample size is fixed. The first principal component contains a rich source of information about high-dimensional two-sample tests. The first eigenvalue is strongly spiked and hence it is difficult to obtain a consistent estimator of the first principle component. In order to overcome the difficulty, we use the noise-reduction methodology by Yata and Aoshima (2012). Ishii et al. (2016, JSPI) thoroughly studied asymptotic properties of the first principal component in HDLSS situations and applied the findings to one sample test and the equality test of two covariance matrices. By using the techniques given by Ishii et al. (2016), we propose a new two-sample test induced by the first principle component. We study asymptotic size and power of the proposed test statistic both theoretically and numerically when the dimension goes to infinity while the sample size is fixed. Finally, we analyze microarray data sets by using several two-sample tests and check the superiority of the proposed test to other candidates.

Modified Variational Mode Decomposition using Ebayesthresh

Guebin Choi

Seoul National University, South Korea. *gbchoi0814@gmail.com*

Abstract: Recently, Dragomiretskiy and Zosso (2014) developed a new decomposition method, termed variational mode decomposition (VMD), which is efficient for handling the problem of tone detection and separation, and for denoising the signal. However, VMD may not be efficient in the presence of missing data since it is based on discrete Fourier transform (DFT) algorithm. To overcome this problem, we propose a new approach based on a novel combination of VMD and Ebayesthresh. The Ebayesthresh provides an effective methodology of getting corrected periodogram. Through simulation study and real data analysis, it is demonstrated that the proposed method can produce substantially effective.

The Kumaraswamy Skew G Distributions

Rui Li

The University of Manchester, United Kingdom. *rui.li-10@postgrad.manchester.ac.uk*

Abstract: Suppose G is an arbitrary cumulative distribution function symmetric around zero. Motivated by Mameli [Statistics and Probability Letters, 104, 2015, 75-81], we introduce a family of skew symmetric generalizations of G, referred to as Kumaraswamy skew G distributions. We derive mathematical properties and estimation issues for the family. The family is capable of producing better fits than all skew symmetric generalizations of the G distribution, as shown by applications to nine real data sets.

On Moment Based Density Approximations for Aggregate Losses

Jeffrey Chu

The University of Manchester, United Kingdom. *jeffrey.chu@postgrad.manchester.ac.uk*

Abstract: Jin et al. (2015) proposed a novel moments based approximation based on the gamma distribution for the compound sum of independent and identical random variables. They illustrated their approximation using six examples. Here, we revisit four of their examples. We show that moments based approximations based on simpler distributions can be good competitors. We also show that the moments based approximations are more accurate than truncated versions of the exact distribution of the compound sum.

Day 4
Thu, June 30

Thu, June 30 (08:30-10:10) | DL01 | Sponsor: IMS

Recent Advances in Machine Learning for Personalized Medicine

Chair: Wenbin Lu (North Carolina State University)

Distinguished Lecturer: Michael Kosorok (The University of North Carolina at Chapel Hill)

Recent Advances in Machine Learning for Personalized Medicine

Michael Kosorok

The University of North Carolina at Chapel Hill, United States. *kosorok@unc.edu*

Abstract: There has recently been an explosion of interest and activity in personalized medicine. However, the goal of personalized medicine—wherein treatments are targeted to take into account patient heterogeneity—has been a focus of medicine for centuries. Precision medicine, on the other hand, is a much more recent refinement which seeks to develop personalized medicine that is empirically based, scientifically rigorous, and reproducible. In this presentation, we describe several new machine learning developments which advance this quest through discovering individualized treatment rules based on patient-level features. Regression and classification are useful statistical tools for estimating such rules based on either observational data or data from a randomized trial, and machine learning approaches can help with this because of their ability to artfully handle high dimensional feature spaces with potentially complex interactions. For the multiple decision setting, reinforcement learning, which is similar to but different from regression, is necessary to properly account for delayed effects. There are several other intriguing nonstandard machine learning tools which can also greatly facilitate discovery of treatment rules. One of these is outcome weighted learning, or O-learning, which directly estimates the decision rules without requiring regression modeling and is thus robust to model misspecification. Several clinical examples illustrating these approaches will also be given.

Tree-based Method for High-dimensional Survival Data

Ruoqing Zhu

University of Illinois at Urbana-Champaign, United States. *teazrq@gmail.com*

Abstract: Tree-based method has been successfully adopted for statistical modeling purposes in personalized medicine. However, its statistical property has not been fully investigated. In this talk, we focus on the consistency of tree-based survival models in high-dimensional setting. A general framework is proposed, which ties the consistency with the choice of splitting rules in the tree construction procedure. The framework includes many existing results as special cases. According to our analysis, we propose a variable importance measure for right censored survival data and construct an associated splitting variable selection rule based on martingale residuals.

How Many Processors Do We Really Need in Parallel Computing?

Guang Cheng

Purdue University, United States. *chengg@stat.purdue.edu*

Abstract: This short talk explores statistical versus computational trade-off to address a basic question in a typical divide-and-conquer setup: what is the minimal computational cost in obtaining statistical optimality? In smoothing spline models, we observe an intriguing phase transition phenomenon for the number of deployed machines that ends up being a simple proxy for computing cost. Specifically, a sharp upper bound for the number of machines is established when the number is below this bound, statistical optimality (in terms of nonparametric estimation or testing) is achievable; otherwise, statistical optimality becomes impossible.

Thu, June 30 (08:30-10:10) | IP26 | Sponsor: Hong Kong

Random Matrices and High-Dimensional Statistics

Organizer: Jianfeng Yao (The University of Hong Kong)

Chair: Debashis Paul (University of California, Davis)

Free Probability and High Dimensional Time Series

Arup Bose

Indian Statistical Institute, India. *bosearu@gmail.com*

Abstract: Large dimensional autocovariance matrices arise naturally in the statistical analysis of high dimensional time series. We study the joint asymptotic behaviour of these matrices in an appropriate sense by establishing connection with free probability. We show how these results can be applied to inference problems such as white noise testing.

More generally, we show the convergence of elements of certain non-commutative space of matrices of increasing dimension generated by appropriate polynomials in non-random matrices and independent random matrices. These are interesting models in free probability in their own right. The resulting limits are described in terms of free circular, free semi-circular and other free variables.

A Universal High-dimensional Data Structural Detection Approach via Random Matrix Theory

Guangming Pan

Nanyang Technological University, Singapore. *gmpan@ntu.edu.sg*

Abstract: We propose to deal with the high-dimensional change point detection problem from a new perspective via random matrix theory. The data dimension p diverges with the sample size n and can be larger than n. Without any specific parametric distribution assumptions and without any estimators, an optimization approach is proposed to figure out both the unknown number of change points and multiple change point positions simultaneously. What's more, an adjustment term is introduced to handle sparse signals when the change only appears in few components out of the p dimension. The computation time is controlled at $O(n^2)$ by adopting a dynamic programming, regardless of the true number of change points k_0. Theoretical results are developed and various simulations are conducted to show the effectiveness of our method. Moreover, as applications, we discuss how to apply the idea proposed in this paper to some other high-dimensional data structure detection problems, e.g. equivalence testing of mean vectors and covariance matrices, which shows the universality of the proposed approach.

Extreme Eigenvalues of Large-dimensional Spiked Fisher Matrices with Application

Qinwen Wang

University of Pennsylvania, United States. *wqw8813@gmail.com*

Abstract: Consider two p-variate populations, not necessarily Gaussian, with covariance matrices Σ_1 and Σ_2, respectively. Let S_1 and S_2 be the corresponding sample covariance matrices with degrees of freedom m and n. When the difference Δ between Σ_1 and Σ_2 is of small rank compared to p, m and n, the Fisher matrix $S := S_2^{-1} S_1$ is called a *spiked Fisher matrix*. When p,m and n grow to infinity proportionally, we establish a phase transition for the extreme eigenvalues of the Fisher matrix: a displacement formula showing that when the eigenvalues of Δ (spikes) are above (or under) a critical value, the associated extreme eigenvalues of S will converge to some point outside the support of the global limit (LSD) of other eigenvalues (become outliers); otherwise, they will converge to the edge points of the LSD. Furthermore, we derive central limit theorems for those outlier eigenvalues of S. The limiting distributions are found to be Gaussian if and only if the corresponding population spike eigenvalues in Δ are simple. Two applications are introduced. The first application uses the largest eigenvalue of the Fisher matrix to test the equality between two high-dimensional covariance matrices, and explicit power function is found under the spiked alternative. The secon application is in the field of signal detection, where an estimator for the number of signals is proposed when the covariance structure of the noise is arbitrary.

Homoscedasticity Tests Valid in Both Low and High-dimensional Regressions

Jianfeng Yao

The University of Hong Kong, Hong Kong. *jeffyao@hku.hk*

Abstract: Testing the homoscedasticity of the errors is a traditional and important problem in a linear regression model. There are in the literature several well-established procedures for this problem such as the White test and the Breusch-Pagan test, and their large-sample theory is also well known. However, in a high-dimensional scenario where the number of covariates p is large compared to the sample size n, these procedures become severely biased. In this paper, we propose two new test statistics to detect the existence of heteroscedasticity. The asymptotic normality of both statistics is obtained under the assumption that the degree of freedom $k=n-p$ tends to infinity. This encompasses in particular two popular settings (i) the classical low-dimensional setting where the number of variables p is fixed while the sample size n tends to infinity; (ii) *the proportional* high-dimensional setting where p and n grow to infinity proportionally such that $p/n \to c \in (0,1)$. This dimension-proof property of the new test procedures guarantees a wide applicability of the proposed procedures to different combinations of the pair (p,n). Extensive Monte-Carlo experiments demonstrate the superiority of our proposed tests over popular existing methods in terms of size and power. The good performance of our tests is also confirmed by several real data analyses. This is a joint work with Zhaoyuan Li.

Thu, June 30 (08:30-10:10) l IP09 l Sponsor: IMS

Adaptive Randomization in Clinical Trials

Organizer: Feifang Hu (The George Washington University)

Chair: Feifang Hu (The George Washington University)

Adaptive Multi-arm Platform Designs for Screening Effective Treatments via Predictive Probability

J. Jack Lee

The University of Texas MD Anderson Cancer Center, United States. *jjlee@mdanderson.org*

Abstract: The process of screening effective treatments one-at-a-time under the current clinical trials system is inefficient and costly. We introduce a statistical framework for designing and conducting randomized multi-arm screening trials using the concept of platform designs and computing Bayesian predictive probability to gauge the treatment effect. The platform designs allow the comparison of multiple arms with a single control arm. It accommodates mid-trial modifications such that poorly performing arms can be dropped early and new agents can be added. When compared to sequential randomized two-arm trials, screening platform designs have the potential to yield considerable reductions in time and cost, eliminate bias stemming from between trial heterogeneity, allowing treating more patients with more effective treatments, and control for multiplicity over a sequence of a priori planned studies, yet, without sacrificing frequentist properties for comparing treatments. The pros and cons of adding response-adaptive randomization will be discussed. Simulations are provided to compare the operating characteristics of the proposed multi-arm platform designs and the sequentially conducted trials.

Joint work with Brian P. Hobbs, Nan Chen

Randomization in Small Population Clinical Trials

William Rosenberger

George Mason University, United States. *wrosenbe@gmu.edu*

Abstract: Rare disease clinical trials present unique problems for the statistician designing the study. First, the trial may be small to reflect the uniquely small population of diseased in the population. Hence, the usual large sample beneficial properties of randomization (balancing on unknown covariates, distribution of standard tests, converging to a target allocation) may not apply. We describe the impact of such trials on consideration of randomization procedures, and discuss randomization as a basis for inference. We conclude that, in small trials, the randomization procedure chosen does matter, and randomization tests should be used as a matter of course due to its property of preserving the type I error rate under time trends.

Central Limit Theorems of a Recursive Stochastic Algorithm with Applications to Adaptive Designs

Li-Xin Zhang

Zhejiang University, China. *stazlx@zju.edu.cn*

Abstract: Stochastic approximation algorithms have been the subject of an enormous body of literature, both theoretical and applied. Recently, Laruelle and Pages (Ann. Appl. Probab., 2013) presented a link between the stochastic approximation and response-adaptive designs in clinical trials based on randomized urn models investigated in Bai and Hu (Ann. Appl. Probab., 2005), and derived the asymptotic normality or central limit theorem for the normalized procedure using a central limit theorem for the stochastic approximation algorithm. However, the classical central limit theorem for the stochastic approximation algorithm does not include all cases of its regression function, creating a gap between the results of Laruelle and Pages (2013) and those of Bai and Hu (2005) for randomized urn models. We establish new central limit theorems of the stochastic approximation algorithm under the popular Lindeberg condition to fill this gap. In our application, we investigate a more involved family of urn models and related adaptive designs in which it is possible to remove the balls from the urn, and the expectation of the total number of balls updated at each stage is not necessary a constant. The asymptotic properties are derived under much less stringent assumptions than those in Bai and Hu (2005) and Laruelle and Pages (2013). The application of the stochastic approximation approach to other types of adaptive designs will also be discussed.

Nonparametric Response Adaptive Randomization Procedures Based on p-values

Zhongqiang Liu

Henan Polytechnic University, China. *zhongqiang@ruc.edu.cn*

Abstract: Many response-adaptive randomization procedures have been proposed and extensively studied in the past decades. However, Most of these procedures are based on parametric structure and do not usually apply to nonparametric situations.

In this paper, we propose a new family of response-adaptive randomization based on p-values of corresponding hypothesis tests. Thus, the proposed procedures can apply to both parametric and nonparametric situations. Under widely satisfied conditions, we derive the asymptotic properties of the procedures and further obtain the power function under nonparametric settings. The proposed procedures are: (i) more robust; and (ii) more ethical than classical response-adaptive randomization procedures under many situations. The advantages are also illustrated in numerical studies.

Thu, June 30 (08:30-10:10) l IP40

Recent Developments in the Analysis of High-Dimensional Time Series with Nonstationarities

Organizer: Haeran Cho (University of Bristol)

Chair: Young K. Lee (Kangwon National University)

Shrinkage Estimation for Multivariate Hidden Markov Models

Mark Fiecas

The University of Warwick, United Kingdom. *m.fiecas@warwick.ac.uk*

Abstract: Motivated from a changing market environment over time, we consider high-dimensional data such as financial returns, generated by a hidden Markov model that allows for switching between different regimes or states. To get more stable estimates of the covariance matrices of the different states, potentially driven by a number of observations which is small compared to the dimension, we modify the EM algorithm to optimise a penalized likelihood. The final algorithm turns out to reproduce better estimates not only for the covariance matrices but also for the transition matrix. It results into a more stable and reliable filter that allows for reconstructing the values of the hidden Markov chain. In addition to a simulation study performed in this paper, we also present a series of theoretical results that includes dimensionality asymptotics and provide the motivation for certain techniques used in the algorithm.

Sparse High-dimensional Varying Coefficient Models

Eun Ryung Lee

Sungkyunkwan University, South Korea. *silverryuee@gmail.com*

Abstract: Varying coefficient models are useful generalizations of parametric linear models. They allow for parameters that depend on a covariate or that develop in time. They have a wide range of applications in time series analysis and regression. In time series analysis they have turned out to be a powerful approach to infer on behavioral and structural changes over time. In this paper, we develop a kernel smoothing approach for sparse high-dimensional varying coefficient models. The estimators make use of a novel penalization scheme working with kernel smoothing. We establish a general and systematic theoretical analysis in high dimensions. Our theory covers regression and time series models. We also address issues of numerical implementation and of data adaptive selection of tuning parameters for penalization. The finite sample performance of the proposed methods is studied by simulations and it is illustrated by an empirical analysis of NASDAQ composite index data.

Non-stationary Dynamic Factor Models for Large Datasets

Matteo Barigozzi

London School of Economics and Political Science, United Kingdom. *m.barigozzi@lse.ac.uk*

Abstract: We develop the econometric theory for Non-Stationary Dynamic Factor models, which allows us to incorporate long-run prediction of macroeconomic theory into the analysis of large datasets. In particular, we derive conditions for consistent estimation of impulse response functions as n and T go to infinity, and for different autoregressive models for the common factors. Finally, the model is used to study the effect of monetary policy shocks on the US economy.

Change-point Detection in High-dimensional Panel Data

Haeran Cho

University of Bristol, United Kingdom. *haeran.cho@bristol.ac.uk*

Abstract: We propose a method for detecting multiple change-points in the mean of (possibly) high-dimensional panel data. CUSUM statistics have been widely adopted for change-point detection in both univariate and multivariate data. For the latter, it is of particular interest to exploit the cross-sectional structure and achieve simultaneous change-point detection across the panel, by searching for change-points from the aggregation of multiple series of CUSUM statistics, each of which is computed on a single series of the panel data. For panel data of high dimensions, the detectability of a change-point is influenced by several factors, such as its sparsity across the panel, the magnitude of jumps at the change-point and the unbalancedness of its location, and having a method that handles a wide range of change-point configurations without any prior knowledge is vital in panel data analysis. The double CUSUM statistic is a determined effort in this direction. We show that the double CUSUM statistic, combined with a binary segmentation algorithm, attains consistent change-point detection in terms of both the total number and the locations of detected change-points, and conduct a comparative simulation study in which its good performance is demonstrated.

Thu, June 30 (08:30-10:10) I IP14 I Sponsor: IISA

Network Models: Theory and Methods

Organizer: Shankar Bhamidi (The University of North Carolina at Chapel Hill)

Chair: Shankar Bhamidi (The University of North Carolina at Chapel Hill)

Geometry of Random Graphs: Scaling Limits and Universality

Sanchayan Sen

TU Eindhoven, Netherlands. *sanchayan.sen1@gmail.com*

Abstract: In the early 2000s, based on non-rigorous arguments, statistical physicists predicted that the typical length of optimal paths in different disordered networks exhibit similar scaling behavior. More precisely, physicists conjectured that for a wide array of random graph models with degree exponent $\tau>3$, distances between typical points both within maximal components in the critical regime as well as on the minimal spanning tree on the giant component in the supercritical regime scale like $n\,\frac{\tau\wedge 4-3}{\tau\wedge 4-1}$. The mathematical machinery available at the time was insufficient for providing a rigorous justification of this conjecture. We report on recent progress in proving this conjecture and characterizing these universality classes in a broader sense.

Dynamic Causal Networks with Multi-scale Temporal Structure

Eric Kolaczyk

Boston University, United States. *kolaczyk@math.bu.edu*

Abstract: I will discuss a novel method to model multivariate time series using dynamic causal networks. This method combines traditional multi-scale modeling and network based neighborhood selection, aiming at capturing the temporally local structure of the data while maintaining the sparsity of the potential interactions. Our multi-scale framework is based on recursive dyadic partitioning, which recursively partitions the temporal axis into finer intervals and allows us to detect local network structural changes at varying temporal resolutions. The dynamic neighborhood selection is achieved through penalized likelihood estimation, where the penalty seeks to limit the number of neighbors used to model the data. Theoretical and numerical results describing the performance of our method will be presented, as well as applications in financial economics and neuroscience. This is joint work with Xinyu Kang and Apratim Ganguly.

Dense and Sparse Graph Limits Arising from Respondent Driven Sampling

Adrian Roellin

National University of Singapore, Singapore. *adrian.roellin@nus.edu.sg*

Abstract: Sampling from large networks is of great statistical importance, since rarely is the full network available to the practitioner. One particular procedure that has been frequently implemented (and often criticised) is the so-called "respondent driven sampling" (also called "snowball sampling"), where participants are asked for referrals, which in turn are again asked for referrals and so on. It has been well-known that this type of sampling will introduce biases, which are determined by the degrees of the nodes in the network. We give a dense and sparse graph limit version of this phenomenon under various implementations of respondent driven sampling. This is joint work with Siva Athreya.

Thu, June 30 (08:30-10:10) | IP36 | Sponsor: Singapore

Statistical Inferences for Complex Data

Organizer: Jin-Ting Zhang (National University of Singapore)

Chair: Gourab Mukherjee (University of Southern California)

High-dimensional Linear Hypothesis Testing Under Heteroscedasticity

Jin-Ting Zhang

National University of Singapore, Singapore. *stazjt@nus.edu.sg*

Abstract: In recent years, with rapid development of data collecting technologies, high-dimensional data become increasingly prevalent. Much work has been done for hypotheses on mean vectors, especially for high-dimensional two-sample problems. Rather than considering a specific problem, we are interested in a general linear hypothesis testing problem on mean vectors of several populations, which include many existing hypotheses about mean vectors as special cases. We propose a test statistic based on a linear combination of U-statistics but it can be quickly calculated without using U-statistics. Asymptotic normality and power of the test are derived under mild conditions without requiring an explicit relation between the data dimension and sample size. Our test is applicable to non-normal multi-sample data without assuming common covariance matrix among different samples. It also works well even when different samples follow different distributions, provided some moment conditions are satisfied. A simple, computation-efficient, and ratio-consistent estimator of the unknown variance of the proposed test statistic is also provided without using computational-intensive U-statistics. Simulation studies and a real data example are presented to demonstrate the good performance of the proposed test.

Modelling Liquidity Supply in Limit Order Book with a Vector Functional Autoregressive (VFAR) model

Ying Chen

National University of Singapore, Singapore. *stacheny@nus.edu.sg*

Abstract: We propose a Vector Functional Autoregressive (VFAR) model to simultaneously describe the dynamics of the liquidity bid and ask supply curves in the limit order book. In particular, we derive a sieve estimator based on the B-splines expansions and the Gaussian process assumption. We investigate the asymptotic properties of the sieve estimators and study its finite sample performance along with simulation study. Based on the bid and ask supply curves of nine stocks traded at the National Association of Securities Dealers Automated Quotations (NASDAQ) stock market in 2015, we show that the VFAR model provides good estimation accuracy and interpretability. This is a joint work with Wee Song Chua.

Principal Flows and Sub-manifolds

Zhigang Yao

National University of Singapore, Singapore. *zhigang.yao@nus.edu.sg*

Abstract: We revisit the problem of extending the notion of principal component analysis (PCA) to multivariate datasets that satisfy nonlinear constraints, therefore lying on Riemannian manifolds. Our aim is to determine curves on the manifold that retain their canonical interpretability as principal components, while at the same time being flexible enough to capture nongeodesic forms of variation. We introduce the concept of a principal flow, a curve on the manifold passing through the mean of the data, and with the property that, at any point of the curve, the tangent velocity vector attempts to fit the first eigenvector of a tangent space PCA locally at that same point, subject to a smoothness constraint. That is, a particle flowing along the principal flow attempts to move along a path of maximal variation of the data, up to smoothness constraints. The rigorous definition of a principal flow is given by means of a Lagrangian variational problem, and its solution is reduced to an ODE problem via the Euler–Lagrange method. Conditions for existence and uniqueness are provided, and an algorithm is outlined for the numerical solution of the problem. Higher order principal flows are also defined. It is shown that global principal flows yield the usual principal components on a Euclidean space. An extension of principal sub-manifolds will be briefly discussed.

(Joint work with Victor Panaretos and Tung Pham)

Density Estimation in the Two-sample Problem with Likelihood Ratio Ordering

Tao Yu

National University of Singapore, Singapore. *stayt@nus.edu.sg*

Abstract: In this paper, we propose a method for estimating the probability density functions in a two sample problem where the ratio of the densities is monotonic. Such a problem is well motivated from medical data and has been widely identified in the literature. However, effective methods for solving this problem are not yet available in the community. Clearly, an effectively method should at least satisfy (1) the resulting estimates are probability densities; (2) the corresponding density ratio inherits the monotonic property, otherwise it is difficult to explain the analysis results. In this paper, we propose estimates for these densities whose ratio inherits the monotonic property, and explore their theoretical properties. One interesting application of our density estimates is that the corresponding receiver operating characteristic (ROC) curve estimate is concave. Through numerical studies, we observe that both the density estimates and the ROC curve estimate from our method outperform their competitors, particularly when the sample size is relatively small.

Thu, June 30 (08:30-10:10) | TCP02

Advanced Modeling of Large-Scale Dependent Data

Organizer: Hiroki Masuda (Kyushu University)

Chair: Hiroki Masuda (Kyushu University)

Statistical Inferences for Ergodic Point Processes and Application to Limit Order Books

Simon Clinet

The University of Tokyo, Japan. *clinet.simon@gmail.com*

Abstract: We construct a general procedure for the Quasi Likelihood Analysis for a multivariate point process on the real half line when its stochastic intensity is ergodic, that is when the law of large numbers applies to functions of the intensity itself along with its first derivatives with respect to the parameters of the model. Under regularity assumptions, we establish the consistency, the asymptotic normality and the convergence of moments of both the Quasi Likelihood estimator and the Quasi Bayesian estimator. In addition, we illustrate our main results by showing how they can be applied to various Limit Order Book models existing in the literature.

Statistical Inference for Price Discovery: a Stochastic Process Approach

Yuta Koike

Tokyo Metropolitan University, Japan. *kyuta@ism.ac.jp*

Abstract: We propose a new framework to investigate price discovery in ultra high frequency trading, which is based on modeling lead-lag effects between martingale components of assets. The framework can be accommodated to non-synchronous trading and market microstructure noise. We show that this framework provides a simple statistical methodology to quantify the price discovery via measuring the lead-lag effects between assets. In particular, we can conduct a test to detect the presence of such a lead-lag effect. The methodology is illustrated by an empirical study to investigate price discovery across different exchanges.

Parametric Inference for Diffusion Processes with High-frequency Financial Data

Teppei Ogihara

The Institute of Statistical Mathematics, Japan. *ogihara@ism.ac.jp*

Abstract: Recently, as availability of intraday security prices data gets increased, the analysis of high-frequency data became more significant. In particular, it is important to estimate security returns' volatility and the covariance of two different securities.

However, when we calculate some statistics using high-frequency data, there are two empirical problems, namely the presence of market microstructure noise and nonsynchronous trading. We propose maximum likelihood type estimation for parametric diffusion processes with discrete and nonsynchronous observations contaminated by market microstructure noise, and prove that our estimator has the best asymptotic variance in any estimators when diffusion processes are Brownian motions. We conjecture our estimator is also the best in general cases.

Mighty Convergence in Mixed-rates Asymptotics

Yusuke Shimizu

Kyushu University, Japan. *y-shimizu@math.kyushu-u.ac.jp*

Abstract: In M-estimation under standard asymptotics, the weak convergence combined with a large deviation estimate of the associated statistical random field provides us with a general tool for deriving not only the asymptotic distribution of the associated M-estimator, but also the convergence of its moments, where the latter plays an important role in theoretical statistics. Here the standard asymptotics refers to the situation where the random field can be well investigated by a single matrix norming, which, however, may be impossible in several situations including sparsely regularized M-estimation. We are concerned here with the uniform tail-probability estimate of a class of scaled M-estimators under mixed-rates asymptotics in the sense of Radchenko (2008), where the associated statistical random fields may fail to be partially locally asymptotically quadratic so that the conventional approach through the polynomial type deviation inequality developed by Yoshida (2011) does not work directly. In particular, our result enables us to deduce convergence of moments of a wide range of regularized M-estimators, which validates, for example, use of AIC-type statistics for tuning-parameter selection in sparse estimation. This is a joint work with Hiroki Masuda.

Thu, June 30 (08:30-10:10) l TCP25

Recent Developments in Large and Complex Data Analysis

Organizer: Xinghao Qiao (London School of Economics and Political Science)

Chair: Shaojun Guo (Renmin University of China)

Nonlinear Shrinkage Estimation of Large Integrated Covariance Matrix

Qilin Hu

London School of Economics and Political Science, United Kingdom. *Q.Hu1@lse.ac.uk*

Abstract: While the use of intra-day price data increases the sample size substantially for asset allocation, the usual realized covariance matrix still suffers from bias contributed from the extreme eigenvalues when the number of assets is large. We introduce a novel nonlinear shrinkage estimator for the integrated volatility matrix which shrinks the extreme eigenvalues of a realized covariance matrix back to acceptable level, and enjoys a certain asymptotic efficiency at the same time, all at a high dimensional setting where the number of assets can have the same order as the number of data points. Compared to a time-variation adjusted realized covariance estimator and the usual realized covariance matrix, our estimator demonstrates favorable performance in both simulations and a real data analysis in portfolio allocation. This include a novel maximum exposure bound and an actual risk bound when our estimator is used in constructing the minimum variance portfolio.

Generalised Additive and Index Models with Shape Constraints

Yining Chen

London School of Economics and Political Science, United Kingdom. *Y.Chen101@lse.ac.uk*

Abstract: We will discuss generalised additive models, with shape restrictions (e.g. monotonicity, convexity, concavity) imposed on each component of the additive prediction function. We show that this framework facilitates a nonparametric estimator of each additive component, obtained by maximising the likelihood. The procedure is free of tuning parameters and under mild conditions is proved to be uniformly consistent on compact intervals. More generally, our methodology can be applied to generalised additive index models. Real data illustrations will be given using our R package 'scar', short for shape-constrained additive regression.

Spatial Weight Matrix Estimation in a Dynamic Spatial Autoregression Model

Cheng Qian

London School of Economics and Political Science, United Kingdom. *C.Qian2@lse.ac.uk*

Abstract: Spatial econometrics focus on cross sectional interaction between physical or economic units. However, most of studies apply a prior knowledge about spatial weight matrix in spatial econometrics model. Therefore misspecification on spatial weight matrix could affect significantly accuracy of model estimation. Lam (2014) has provided an error upper bound for the spatial regression parameter estimators in a spatial autoregression model, showing that misspecification can indeed introduce large bias in the final estimates.

Meanwhile, new researches on spatial weight matrix estimation only consider static effects but not include dynamic effects between spatial units. Our model firstly use the different linear combinations of same spatial weight matrix specifications for different time-lag responds in proposed spatial econometrics model. Moreover, by introducing penalization on the coefficients of the linear combination of spatial weight matrix specifications, the best specification or the best linear combination of specifications for different lag spatial effect can be selected. To overcome endogeneity from autoregression, instrumental variables are introduced. The model we use in this paper can also find fixed effects and spillover effects. Finally, we also develop asymptotic normality for our estimation under the framework of functional dependence measure introduced in Wu (2011).

Multi-zoom Autoregressive Models

Rafal Baranowski

London School of Economics and Political Science, United Kingdom. *R.Baranowski@lse.ac.uk*

Abstract: Let X_t be a univariate time series representing returns on a financial instrument, observed at a mid- or high-frequency, e.g. one-minute. We consider the problem of statistical modelling of X_t using its own past values, with a focus on prediction. This is a notoriously difficult task, as the financial returns are typically very weakly autocorrelated, hence resemble a white-noise-like process. To address this issue, Fryzlewicz (2013) introduced so-called Multi-Zoom Autoregressive (MZAR) time series models, where X_t depends on a few past returns observed at lower scales, such as one hour or one day. Due to the multi-scale structure, MZAR models mimic white noise from the point of view of the sample autocorrelation, hence are potentially useful in analysis of the financial returns.

We propose an estimation procedure for fitting MZAR models to the data. The procedure consists of fitting of a large AR model via OLS and subsequent change-point analysis using Wild Binary Segmentation methodology performed on the estimated coefficients in order to identify the relevant time scales. In the final step, OLS is used again to fit the model with the selected scales. We prove that this procedure is consistent in a high-dimensional framework, where the maximum lag of X_t included in the model diverges with n. We provide recommendations on the default choices of the parameters for our procedure. In an empirical analysis of the data from the New York Stock Exchange Trades and Quotes database, we demonstrate that the fitted MZAR models offer very good performance in predicting high- and mid- frequency returns.

This is a joint work with Prof. Piotr Fryzlewicz.

Thu, June 30 (08:30-10:10) l TCP34

Statistical Modeling and Its Applications

Organizer: Xuejun Jiang (South University of Science and Technology of China)

Chair: Xuejun Jiang (South University of Science and Technology of China)

Detection of Gene Regulatory Relationships by ODE Model with Time Lags

Jie Hu

Xiamen University, China. *hujiechelsea@hotmail.com*

Abstract: Genes play one of the most important and essential roles in regulation and control of organism's varieties of phenotype and behavior. Hence, clarifying the regulation mechanism of the expression of genes is the key step to control the phenotype of organism. By searching regulatory relationships among genes, we try to detect the mysterious and important gene regulatory network. However, traditional models cannot capture the regulatory mechanism precisely. Based on gene expression data sets measured in yeast cells and Ordinary Differential Equation (ODE) model, we now aim at developing a probabilistic framework for evaluating possible regulatory relationships among profiled genes. Regulatory models with one or two parents are considered. Time series of the measured genes are fitted into a dynamics model based on time-lagged regression. An information-criterion type of measure is used to select statistically significant gene regulatory relationships. Known regulatory relationships in current literatures and databases are used to evaluate our results.

An Extended Mallows' Model for Rank Data Aggregation with Covariates

Han Li

Shenzhen University, China. *hli@szu.edu.cn*

Abstract: By extending the Mallows' model, we propose a new model for rank data that incorporates parameters for overall as well as position dependent reliability measure. The ranking model has closed form expressions for both the full ranking and the partial ranking lists, thus it provides a flexible framework for statistical inference in practice. Besides, adopting the LASSO approach, our model could select the covariates that affect the ranking of the items. Finally we apply the proposed model to extensive synthetic datasets and real examples to evaluate its merits.

Spatial Analysis of Water Quality in Fujian Bay, China

Yan Liu

Ocean University of China, China. *liuyan_ouc@126.com*

Abstract: The coastal environment is dynamic, complex and site-specific, and greatly disturbed by human activities. In this study, we try to analyze the seawater quality trends of Fujian Province, China. The seawater quality parameters chosen are pH, dissolved oxygen (DO), chemical oxygen demand (COD), dissolved inorganic nitrogen (DIN, including NO2-N, NO3-N and NH4-N), phosphate (PO4-P) and petroleum hydrocarbons (Oil). We collect the quarterly data of 103 monitoring sites from 13 bays, 2007-2014. Seasonal trends are studied via Mann-Kendall's test. Cluster analysis and spatial analysis combining oceanographic knowledge are conducted to study the spatial variations and interactions of seawater parameters. Finally, the relationship between bay environment and human activities are established.

The Latent Low Rank Model to Colocalize Genetic Risk Variants in Multiple GWAS

Jin Liu

Duke-NUS Medical School, Singapore. *jin.liu@duke-nus.edu.sg*

Abstract: To date, more than one thousand genome-wide association studies (GWAS) have been completely. Due to the polygenic architecture of complex diseases, identification of risk single-nucleotide polymorphism (SNP) markers remains challenging. Until recently, most of conventional statistical methods only investigate one GWAS data set of one trait/disease at a time. To model the genetic correlation among complex diseases (formally known as "pleiotropy"), GPA [Chung et al., 2014] and EPS used "four-group model"; for two GWAS. The model complexity of parameter space grows exponentially as the number of GWAS increases. Here, we proposed LLR, the Latent Low Rank model to jointly analyze multiple GWAS. LLR uses a latent variable Z to indicate the null and non-null states as a "two-group model"; for each trait meanwhile the prior probability of the latent variable Z is modulated by a low-rank matrix X which is constrained by the nuclear norm. To estimate parameters in the corresponding penalized complete-data log-likelihood of LLR, we showed the penalized EM algorithm based on EM and accelerated proximal gradient (APG) algorithms as the benchmark. To improve its efficiency, we developed an extremely efficient path algorithm - EM boosting. EM boosting algorithm updates parameters in the M-step using the forward stagewise procedure [Tibshirani, 2014] and is capable of building up a whole path efficiently. We compared the similarity of the two algorithms using a synthetic dataset and used synthetic datasets to show that LLR greatly improved the power of identification of risk markers by extensive simulation study. We applied LLR to jointly analyze p-values for 19 traits.

Thu, June 30 (08:30-10:10) | TCP35

Statistical Modeling in Economics and Finance

Organizer: Yan Liu (Ocean University of China)

Chair: Xin Zhao (Ocean University of China)

Energy Consumption and Economic Growth: Evidence from Coastal Areas in China

Jing Guo

Ocean University of China, China. *oucguojing@163.com*

Abstract: This study examines the relationship between energy consumption and economic growth in China's coastal areas, analyzing capital and labor as potential determinants of production function. We adopt semiparametric regressions for the three marine economic zones (North, East and South) over the period between 1990 and 2013. This empirical evidence supports a unidirectional causality running from energy consumption to economic growth, among which the South Sea Zone shows the highest level of dependence on energy consumption. However, the first-order derivatives of nonparametric estimates reveal that over time the North and East Sea Zone suffer from inefficient energy use. Our analysis informs policy makers about improving energy efficiency that would ensure sustainable economic development in China's coastal areas.

Empirical Research on the Market Volatility of Copper Future of SHFE

Ruifen Huang

Ocean University of China, China. *qdhdhrf@163.com*

Abstract: Futures market is a critical part of the capital market. As the most important trading type of SHFE copper future is the most active and one of the largest trading volume of futures contracts. So accurate understanding of the inherent law of the futures market's price fluctuation can help identify risk. This paper analyses the fluctuation of futures market through fat-tail distribution, volatility clustering, leverage effect and international linkage of SHFE using GARCH, TGARCH, GJR-GARCH and VECM model with daily closing price of copper futures from March 6th 2009 to November 5th 2015. The study of these characteristics can provide reference for the regulators who will make new policies and help investors make right investment strategies.

Research on Marine Economy Efficiency and Its Influencing Factors in the Blue Economic Region of China

Yong Peng

Ocean University of China, China. *819498094@qq.com*

Abstract: Under the background of "Ocean Power Strategy", the position of marine economy in the national economy is becoming more and more important. In order to make clear the level of marine economic efficiency and its influencing factors in the blue economic region of China, this paper uses SFA (Stochastic Frontier Analysis) to measure marine economic efficiency, uses ESDA(Exploratory Spatial Data Analysis) to explore its temporal and spatial distribution pattern, uses Spatial Panel Econometric Models to analysis its influencing factors. The data of this paper is panel data for China's eleven coastal provinces from 2008 to 2013. The results show that most coastal provinces' marine economic efficiency are in the middle level, efficiency difference is becoming smaller and smaller between different regions, there is spatial spillover effect between adjacent regions, and industrial structure is a key factor to influence the marine economic efficiency. Moreover, our research results could provide reference for decision makers to develop marine economy.

What Matters More? Developing an Integrated Weighting Technique for Coastal Vulnerability to Storm Surge

Shun Yuan

Ocean University of China, China. *yshydx@163.com*

Abstract: Weighting different indicators plays a critical role in quantifying vulnerability to natural disasters. However, there is no consensus about which weighting method performs best. This paper proposes a method to aggregate different weighting techniques to quantify vulnerability to storm surge using social, economic and environmental indicators. Viewing vulnerability through exposure, sensitivity, and adaptability dimensions, we used a geographic information system to establish areas of vulnerability for China's eleven coastal provinces. Our results suggest that vulnerability to storm surge varies drastically from one area to another, yet does not follow expected geographic boundaries. Shandong province is the most vulnerable area while Shanghai is the least. Subgroup weights show that exposure and sensitivity drive vulnerability. Our analysis enables policy makers to identify an area's strength and weakness in disaster management, and integrate these criteria into more robust vulnerability evaluation.

Does the Relationship Between Insurance Development and Economic Growth Maintain Stability? An Empirical Analysis of Coastal Area in China Based on Non-parametric Local Polynomial Regression

Hui Zheng

Ocean University of China, China. *qdzhouc@163.com*

Abstract: This paper sets the relationship between insurance development and economic growth as its research foundation, with the Northern, Eastern and Southern coastal areas being the research object. Based on non-parametric local polynomial regression, it studied the dynamic change between insurance development and economic growth. It obtained a curve fitting of insurance density and GDP per capita and analyzed fitting effects and the monotonicity, convexity and dynamic change characteristics of the curve. The results demonstrated that there were obvious differences in output efficiency and marginal revenue between the three regions, and around 2003, the relationship between the two variables changed from complex to simple. With the fitting line of the whole area as the reference standard, the relationship in the Northern and Southern regions was stable, but the curve changed complicatedly in the Eastern area. As a whole, it was certain that insurance growth had a positive effect on the economic development of the coastal area, and the law of diminishing returns held in most cases. Those conclusions offered several counter measures for policy-makers and researchers.

Thu, June 30 (08:30-10:10) | CP09

Probability Session 2

Chair: Phillip Sheung Chi Yam (The Chinese University of Hong Kong)

Favourable Non-extreme Region: Outlier

Tudzla Hernita

STIS 53 Computer Division, Indonesia. *hernitatudzla@gmail.com*

Abstract: Usual connotation of outlier is extreme. This connotation can be true for univariate. For bivariate another term is introduced that is leverage. It is safe to say that extreme point can leverage. It is also safe to say that outlier which happen to be extreme can be called leverage. Firstly univariate leverage is excluded. Secondly bivariate outlier is identified which happen to be non-extreme by means of Minimum Covariance Determinant (MCD). International comparison of iPad price is used as an introduction to intra Indonesia comparison of Human Development Index (HDI) by iterative process. The research goal is to apply different treatment for identified non-extreme region (outlier). Using bivariate 2014 data of per caput Gross Domestic Product (GDP) and iPad price from 30 countries after Singapore and Brazil leverage are separated, iteratively a country is taken out and MCD is calculated 30 times. Further iteratively two countries are taken out at a time and MCD is calculated 435 times. Both GDP and iPad price have highest level of measurement as required by MCD calculation. Arab Emirates is a non-extreme outlier having relatively high GDP with relatively low iPad price. This favourable case may promote possession of iPad. In Indonesia human development applies to all regions, however certain regions may worth a development acceleration. Thirdly MCD is applied to 434 regions for bivariate HDI and GDP data excluding Bojonegoro as leverage. Among 434 regions having GDP under ten trillion Rupiah and HDI over sixty, Probolinggo is different that is within the vicinity of regions having similar GDP around 7,5 Probolinggo has a lower HDI and simultaneously within the vicinity of regions having similar HDI around 64,5 Probolinggo has a higher GDP. Probolinggo is not a bivariate extreme but a bivariate outlier. Government acceleration of human development can start with identified non-extreme outlier.

Randomly Weighted Sums of Conditionally Dependent Random Variables with Applications to Risk Theory

Dongya Cheng

Soochow University, China. *dycheng@suda.edu.cn*

Abstract: Consider the randomly weighted sums and their maximum, where the primary random variables are real-valued with a common subexponential distribution, and the random weights are nonnegative and independent of the primary random variables. Under some conditional dependence assumptions on the primary random variables, we investigate tail behavior of the randomly weighted sums and their maximum. The obtained result is applied to estimating finite-time ruin probabilities in a discrete-time risk model with both insurance and financial risks.

Precise Local Large Deviations for Random Sums with Applications to Risk Models

Fengyang Cheng

Soochow University, China. *chengfy@suda.edu.cn*

Abstract: In this paper, we investigate the precise local large deviation probabilities for random sums of independent real-valued random variables with a common distribution F, where $F(x+\Delta)=F((x,x+T])$ is an O-regularly varying function for some fixed constant $T>0$(finite or infinite). We also obtain some results on precise local large deviation probabilities for the claim surplus process of generalized risk models in which the premium income until time t is simply assumed to be a nondecreasing and nonnegative stochastic process. In particular, the results we obtained are also valid for the global case, i.e. case $T=\infty$.

Kolmogorov-type Inequality and Some Limit Theorems for Extended Negatively Dependent Random Variables

Jigao Yan

Soochow University, China. *yanjigao@suda.edu.cn*

Abstract: In this paper, we give the Kolmogorov-type inequality and some limit theorems (SLLN and Three Series Theorem, etc.) for Extended Negatively Dependent(END) random variables. As the application, we also consider the case of Jamison weighted sums, which extend the corresponding results for NA random variables.

Thu, June 30 (10:30-12:10) l DL05 l Sponsor: IMS

Particle Representations for Measure-Valued Processes and Stochastic Partial Differential Equations

Chair: Sunder Sethuraman (University of Arizona)

Distinguished Lecturer: Thomas G. Kurtz (University of Wisconsin-Madison)

Particle Representations for Stochastic Partial Differential Equations

Thomas G. Kurtz

University of Wisconsin-Madison, United States. *kurtz@math.wisc.edu*

Abstract: Stochastic partial differential equations arise naturally as limits of finite systems of weighted interacting particles. For a variety of purposes, it is useful to keep the particles in the limit obtaining an infinite exchangeable system of stochastic differential equations for the particle locations and weights. The corresponding de Finetti measure then gives the solution of the SPDE. These representations frequently simplify existence, uniqueness and convergence results.

Beginning with the classical McKean-Vlasov limit, the basic results on exchangeable systems along with several examples will be discussed.

SPDE with Poisson Representation

Jie Xiong

University of Macau, Macao. *jiexiong@umac.mo*

Abstract: Representation of the solution to stochastic partial differential equations (SPDE) by particle system with branching and time-varying weight has been studied extensively. It has applications in providing numerical approximation to the solution of SPDE, especially, the filtering equation. In this talk, we will present another type of representation for the solution of SPDEs. We will demonstrate the advantages of this representation. The uniqueness of the solution to this representation is obtained by a large deviation type estimation. This talk is based on a joint research with Tom Kurtz.

Uniform in Time Particle System Approximations for Nonlinear Equations of Keller-Segel Type

Wai Tong Fan

University of Wisconsin-Madison, United States. *ariesfanhk@gmail.com*

Abstract: We consider a collection of diffusing particles that interact indirectly through a dynamic chemical field produced by the particles themselves. The long time dynamics of this system is poorly understood, yet simulations suggest interesting aggregation behavior. In this talk, I will present a uniform in time propagation of chaos for the particle system, which implies the law of large numbers for the empirical measures. The limit is characterized through a nonlinear PDE of the Keller-Segel type. Uniform in time exponential concentration bounds and results on uniform in time convergence of an Euler approximation scheme will also be presented. This is joint work with Amarjit Budhiraja.

Thu, June 30 (10:30-12:10) l IP45

Random Fields: Theory and Applications

Organizer: Yimin Xiao (Michigan State University)

Chair: Zhengjun Zhang (University of Wisconsin-Madison)

Local Times of Gaussian Random Fields on the Sphere

Xiaohong Lan

University of Science and Technology of China, China. *xhlan@ustc.edu.cn*

Abstract: We shall be concerned with the property of strong local nondeterminism for Gaussian random fields T(x) on the sphere and apply it to study the existence, joint continuity and Hölder regularity of the local times of T(x). These results can be exploited to study analytic and geometric properties of the sample path of spherical random fields..

Gaussian Random Fields with Stationary Increments and their Asymptotic Properties

Wensheng Wang

Hangzhou Normal University, China. *wswang@stat.ecnu.edu.cn*

Abstract: Let $\{Y(t), t \in R^N\}$ be a real-valued Gaussian random field with stationary increments and $Y(0)=0$. Consider the (N,d)-Gaussian random field defined by $X(t)=(X_1(t),\ldots,X_d(t))$, where $X_1,\ldots,X_d$ are independent copies of Y. The moduli of non-differentiability and Strassen's laws of the iterated logarithm in the set variable for sample paths of Gaussian random fields with stationary increments are established and the regular properties of the local time of $X(t)$ are established.

These results give precise information about the differentiability and the global and local maximum fluctuations of the sample functions. Applications to fractional Riesz-Bessel processes and the Cauchy class are considered.

Wavelet Estimators of Multivariable Nonparametric Regression Functions with Long Memory Data

Dongsheng Wu

The University of Alabama in Huntsville, United States. *dongsheng.wu@uah.edu*

Abstract: In this talk, we propose wavelet estimators for multivariable nonparametric regression functions with long memory data and investigate the asymptotic rates of convergence of wavelet estimators based on level dependent thresholding and block thresholding, respectively. For the level dependent thresholding, we provide an asymptotic expansion for the mean integrated squared error (MISE) of the estimators. For the block thresholding, we show that the estimators achieve optimal minimax convergence rates over a large class of functions. These results extend the corresponding results of Li and Xiao (2006, 2007) to the multivariable setting.

This talk is based on joint works with Yunzhu He and Yimin Xiao.

Estimation of Fractal Indices for Bivariate Gaussian Random Fields

Yimin Xiao

Michigan State University, United States. *xiao@stt.msu.edu*

Abstract: This talk is concerned with estimation of fractal indices of multivariate Gaussian random fields (or spatial processes). In the multivariate setting, it is important and challenging to quantify the effect of the cross dependence structure on the joint performance of the estimators. By extending the increment-based method (Chan and Wood, 2000, 2004; Anderes, 2010), we construct estimators for the fractal indices of a large class of locally stationary bivariate Gaussian random fields and investigated their joint statistical performance. We provide conditions on the cross covariance for the estimators to be asymptotically independent. The main results are applicable to the bivariate stationary Gaussian random fields with Mat´ern cross covariance functions introduced by Kleiber, Gneiting and Schlather (JASA, 2010).

(This talk is based on joint work with Yuzhen Zhou.)

Thu, June 30 (10:30-12:10) I IP49

Emerging Statistical Methods in Big Data Analytics

Organizer: Ping Ma (University of Georgia)

Chair: Hongkai Ji (Johns Hopkins University)

Pooling Partial Observations for Efficient Estimation of the Joint Distribution

Xiaodan Fan

The Chinese University of Hong Kong, Hong Kong. *xiaodan.fan@gmail.com*

Abstract: One aspect of high-dimensional big data is that many observations are partial in the sense that only a small part of the interested variables are observed for each observation. To make predictions based on these data, we often need the full picture of the joint distribution of all interested variables, which is hidden in these partial observations. We developed a robust and efficient method for constructing the joint distribution efficiently from Boolean big data with high missing rate. The method is based on Dirichlet process mixture of product multinomials. Both synthetic data and real data showed the high accuracy of our method.

Automated Feature Identification using Online Knowledge and EMR Data

Sheng Yu

Tsinghua University, China. *yusir@outlook.com*

Abstract: Clinical, pharmaceutical, and genetic association studies usually require patients with a specific disease to be identified. Hospitals with electronic medical record (EMR) systems, especially the large and early adopters, have accumulated enormous healthcare data, from which target people can be potentially identified in large numbers with so-called phenotyping algorithms. To develop such an algorithm, natural language processing (NLP) is needed to extract information from the clinical notes, and machine learning can utilize the NLP data and other information to classify if a patient is a target. However, in the conventional work flow, medical concepts and terms related to the disease, and features for the machine learning, need to be curated by medical experts of the domain, with the assistance of medical informaticians and statisticians, which is a time consuming and expensive process.

To reduce the reliance on experts and achieve large scale development of phenotyping algorithms, we developed a method and system that leverage the online medical knowledge sources and EMR data to automatically curate medical concepts and features. Medical concepts related to the target disease are identified from publicly available knowledge sources such as the Wikipedia with NLP. Association patterns of the identified concepts in the EMR database are then used to select the most informative ones as features for machine learning. Validations using the gold-standard training sets from previous studies have shown that accuracy of algorithms trained with the automatically curated features rivals those trained with expert-curated features. This method substantially reduces the time and cost of phenotyping algorithm development, and it has been employed for building large scale biobanks in major medical research institutes in the U.S.

Discovering Association Patterns via Theme Dictionary Models

Ke Deng

Tsinghua University, China. *kdeng@math.tsinghua.edu.cn*

Abstract: Discovering patterns from a set of text or, more generally, categorical data is an important problem in many disciplines such as biomedical research, linguistics, artificial intelligence and sociology. We consider here the well-known 'market basket'problem that is often discussed in the data mining community, and is also quite ubiquitous in biomedical research. The data under consideration are a set of 'baskets', where each basket contains a list of 'items'. Our goal is to discover 'themes', which are defined as subsets of items that tend to co-occur in a basket. We describe a generative model, i.e. the theme dictionary model, for such data structures and describe two likelihood-based methods to infer themes that are hidden in a collection of baskets. We also propose a novel sequential Monte Carlo method to overcome computational challenge. Using both simulation studies and real applications, we demonstrate that the new approach proposed is significantly more powerful than existing methods, such as association rule mining and topic modeling, in detecting weak and subtle interactions in the data.

Impact of Genotyping Errors on Statistical Power of Association Tests in Genomic Analyses

Lin Hou

Tsinghua University, China. *houl@tsinghua.edu.cn*

Abstract: A key step in genomic studies is to assess high throughput measurements across millions of markers for each participant's DNA, either using microarrays or sequencing techniques. Accurate genotype calling is essential for downstream statistical analysis of genotype-phenotype associations. In addition, next generation sequencing (NGS) has recently become more a common approach in genomic studies. How the accuracy of variant calling in NGS-based studies affects downstream association analysis has not, however, been studied using empirical data in which both microarrays and NGS were available. In this article, we investigate the impact of variant calling errors on the statistical power to identify associations between single nucleotides and disease, and on associations between multiple rare variants and disease. Our results show that the power of burden test for rare variants is strongly influenced by the specificity in variant calling, but is rather robust with regard to sensitivity. By using the variant calling accuracies estimated from a sub-study of a Cooperative Studies Program project conducted by the Department of Veterans Affairs, we show that the power of association tests are mostly retained with commonly adopted variant calling pipelines.

Thu, June 30 (10:30-12:10) I IP47

Emerging Nonparametric Methods for Financial Data

Organizer: Lan Xue (Oregon State University)

Chair: Fang Li (Indiana University-Purdue University Indianapolis)

High-dimensional Inference with an Application to Financial Networks

Yongli Zhang

University of Oregon, United States. *yongli@uoregon.edu*

Abstract: Analysis of a network's structure has become increasingly important in understanding systemic risk inherited in a financial system. In this article, we studies high-dimensional inference for graphical models, where undirected graphs have been reconstructed and explored for network analysis, as in financial networks, describing pairwise associations among a large number of interacting units. In particular, we develop a bootstrap method to account for variation inherited in regularization. Moreover, inference concerning a network's structure is formulated in hypothesis testing in the presence of nuisance parameters, and our inference procedure for linkage, the parameter of interest, does not regularize it. This is in contrast to common practice that such an inference could be biased depending on if it is regularized to be "zero", which is an issue commonly encountered in inference after model selection. In theory, we show that the method leads to correct asymptotic inference in a high-dimensional situation. Numerically, it compares favorably against existing methods. An application to inferring a financial network's structure involving 200 publicly traded stocks is described and the effect of Lehman's collapse on the network's topology is revealed.

Spline Confidence Bands for Generalized Regression Models

Jing Wang

University of Illinois at Chicago, United States. *jiwang12@uic.edu*

Abstract: A computational study of bootstrap confidence bands based on a free-knot spline regression is explored for the generalized linear models in this paper. In free-knot spline regression, the knot locations as additional parameters offers greater flexibility and the potential to better account for rapid shifts in slope and other important structures in the target function. However, the search for optimal solutions becomes very complicated because of "freeing" up the knots. In particular, the lethargy" property in the objective function results in many local optima with replicate knot solutions. To prevent solutions with identical knots, a penalized Quasi-likelihood estimating equation is proposed that relies on both a Jupp transformation of knot locations and an added penalty on solutions with small minimal distances between knots. Focusing on logistic regression for binary outcome data, a parametric bootstrap is used to study the variability of the proposed estimator and to construct confidence bands for the unknown form of the logistic regression link function. This is a joint work with Dr. Ella Revzin at Coyote Company at Chicago.

Monotone Additive Models in Productivity Analysis

Lan Xue

Oregon State University, United States. *xuel@stat.oregonstate.edu*

Abstract: Monotone additive models are useful in estimating productivity curve or analyzing disease risk where the predictors are known to have monotonic effects on the response. Existing literature mainly focuses on univariate monotone smoothing. Available methods for estimation of monotone additive models are either difficult to interpret or have no asymptotic guarantees. In this paper, we propose a one-step backfitted constrained polynomial spline method for monotone additive models. It is not only easy to compute by taking numerical advantages of linear programming, but also enjoys the optimal rate of convergence asymptotically. The simulation study and application of our method to Norwegian Farm data suggest that the proposed method has superior performance than the existing ones, especially when the data has outliers.

Oracally Efficient Estimation and Consistent Model Selection for ARMA Time Series with Trend

Lijian Yang

Soochow University, China. *yanglijian@suda.edu.cn*

Abstract: Most time series encountered in practice contain nonzero trend, yet textbook approaches to time series analysis are typically focused on zero mean stationary autoregressive moving-average (ARMA) processes. Trend is often estimated by ad hoc methods and subtracted from time series, and the residuals are used as the true ARMA noises for data analysis and inference, including parameter estimation, lag selection and prediction. We propose a theoretically justified two-step method to analyze time series consisting of a smooth trend function and ARMA error term, which is computationally efficient and easy for practitioners to implement. The trend is estimated by B-spline regression, and the maximum likelihood estimator (MLE) based on residuals is shown to be oracally efficient in the sense that it is asymptotically as efficient as if the true trend function were known and then removed so as to obtain the ARMA errors. In addition, consistency of the Bayesian information criterion (BIC) for model selection is established for the detrended residual sequence. Finite sample performance of the proposed procedure is illustrated by simulation studies and real data analysis.

Thu, June 30 (10:30-12:10) | IP59

Advances in Analysis of Complex High-Dimensional Data

Organizer: Shaojun Guo (Renmin University of China)

Chair: Xin Tong (University of Southern California)

Estimation of Nonsmooth Functionals

Mark Low

University of Pennsylvania, United States. *lowm@wharton.upenn.edu*

Abstract: In this talk I will discuss some estimation problems where standard approaches fail. These problems exhibit some interesting features that are significantly different from those that occur in estimating smooth functionals. I will discuss a general lower bound technique and illustrate the ideas by focusing on optimal estimation of the L^1 norm.

Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity

Tony Cai

University of Pennsylvania, United States. *tcai@wharton.upenn.edu*

Abstract: Confidence sets play a fundamental role in statistical inference. In this paper, we consider confidence intervals for high dimensional linear regression with random design. We first establish the convergence rates of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter is given. The focus is then on the problem of adaptation to sparsity for the construction of confidence intervals. Ideally, an adaptive confidence interval should have its length automatically adjusted to the sparsity of the unknown regression vector, while maintaining a prespecified coverage probability. It is shown that such a goal is in general not attainable, except when the sparsity parameter is restricted to a small region over which the confidence intervals have the optimal length of the usual parametric rate. It is further demonstrated that the lack of adaptivity is not due to the conservativeness of the minimax framework, but is fundamentally caused by the difficulty of learning the bias accurately.

Valid Inference on Semiparametric Estimators with Regressors Generated by High Dimensional Regularization

Shaojun Guo

Renmin University of China, China. *guoshaoj@amss.ac.cn*

Abstract: In many situations, estimation on the parameter of interest often involves covariates which are not directly observable but could be estimated from data in the first step. These so-called generated covariates appear in numerous applications, including two-stage nonparametric regression and censored regression models. In this paper, we focus on the problem where covariates are generated through high dimensional regularization. It turns out that the regularization step has a very serious effect for valid inference on parameters of interest. Our primary interest is to develop a novel regularized approach to generate covariates. The proposed estimator can be shown to be asymptotically normal. To illustrate, we provide several examples to demonstrate the superiority of the proposed approach. This approach is also applicable to linear or nonlinear functionals in other sparse nonparametric high dimensional regression models such as additive or varying coefficient models.

High Dimensional and Functional Autoregressions with an Application to Functional Volatility Processes

Xinghao Qiao

London School of Economics and Political Science, United Kingdom. *X.Qiao@lse.ac.uk*

Abstract: Modelling multiple curves arises in a broad spectrum of real applications. However, many studies in functional data literature focus primarily on the critical assumption of independent and identically distributed (i.i.d.) samples. In this talk, we proposed vector-valued functional autoregressive models (VFAR) to characterize the temporal and cross-sectional dependence structure across high dimensional curve time series. We develop a regularization approach via the group lasso penalty to estimate the autoregressive coefficient functions. We also introduce a functional stability measure for stationary functional processes that provides insight into the effect of dependence on the accuracy of regularized estimates and derive the non-asymptotic bounds for the estimation errors of the regularized estimates. Finally, we show that the proposed methodology significantly outperforms its competitors through a series of simulations and one high-frequency daily trading dataset.

Thu, June 30 (10:30-12:10) l TCP09

Recent Advances in Biomarker Evaluation and Risk Prediction

Organizer: Shanshan Li (Indiana University)

Chair: Ruoqing Zhu (University of Illinois at Urbana-Champaign)

Variable Selection and Predictive Performance for Correlated Biomarkers using Regularized Regression Approaches

Wei-Ting Hwang

University of Pennsylvania, United States. *whwang@mail.med.upenn.edu*

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 useful subset of biomarkers from a group of high-dimensional candidate biomarkers. However, these variable selection methods can give unstable results through cross-validation process and may be influenced by the correlations between candidate biomarkers and other factors. 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, size of candidate biomarkers, presence of confounders, etc.) on signature composition. We proposed a simple solution to address the problem of stability and an alternative nested cross-validation process in selecting tuning parameters values. 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.

Estimation of Covariate-Specific Time-Dependent ROC Curves in the Presence of Missing Biomarkers

Shanshan Li

Indiana University, United States. *sl50@iu.edu*

Abstract: Covariate-specific time-dependent ROC curves are often used to evaluate the diagnostic accuracy of a biomarker with time-to-event outcomes, when certain covariates have an impact on the test accuracy. In many medical studies, measurements of biomarkers are subject to missingness due to high cost or limitation of technology. In this talk, we consider estimation of covariate-specific time-dependent ROC curves in the presence of missing biomarkers. To incorporate the covariate effect, we assume a proportional hazards model for the failure time given the biomarker and the covariates, and a semiparametric location model for the biomarker given the covariates. To deal with the missing biomarkers, we propose a simple weighted estimator for the ROC curves where the weights are inversely proportional to the selection probability. We also propose an augmented weighted estimator which utilizes information from the subjects with missing biomarkers. We derive the large sample properties of the proposed estimators and evaluate their finite sample performance using numerical studies. The proposed approaches are illustrated using the US Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

Personalized Differential Gene Expression Detection

Yingying Wei

The Chinese University of Hong Kong, Hong Kong. *yweicuhk@gmail.com*

Abstract: Standard differential gene expression (DE) detection methods compare gene expression level of all the cases versus all the controls. In contrast, outlier analysis approaches such as COPA and outlier sums based methods focuses on abnormal gene expression in only a minority of samples. Nevertheless, outlier analysis still calls DE at gene level for the whole cohort. In this study, we propose an empirical Bayes method to call DE for each individual patient, thus enabling personalized biomarker prediction.

Combine Longitudinal Biomarkers in Predicting Time-dependent Risks

Hong Zhu

University of Texas Southwestern Medical Center, United States. *hong.zhu@utsouthwestern.edu*

Abstract: In many medical practice settings, patients are monitored sequentially and biomarkers are repeated collected on the same patient over time with the goal of updating prognosis and correctly classifying patients with different predicted risk profiles. Recent technological developments in genomic and biomedical research have generated a large number of candidate longitudinal biomarkers (e.g., protein expression, RNA expression or single nucleotide polymorphisms) that have potential to be used in the prediction of clinical outcome. Combining longitudinal biomarkers often substantially improves the predictive accuracy over baseline markers and statistical methods are needed to evaluate the predictive accuracy associated with longitudinal biomarkers. A major challenge is that the predictive accuracy may be different depending on the time when the biomarker is measured, because markers are measured longitudinally and usually vary with time. Another challenge arises when the primary outcome is the time to an event. The disease status of subjects can change over time and the disease onset time may be censored. The concept of receiver operating characteristic (ROC) curve has been extended to integrate the time dimension in the analysis. Particularly, developing methods for assessing the predictive accuracy of longitudinal biomarkers has received extensive research attention (Slate and Turnbull, 2000; Zheng and Heagerty, 2004). We propose a robust method to combine longitudinal biomarkers in predicting time-dependent risks by directly modeling the area under the ROC curve (AUC). The proposed method that incorporates the measurement time would allow for an updated medical decision at different subsequent time points over the follow-up period. It is also robust to model mis-specification for survival time and marker distributions.

Thu, June 30 (10:30-12:10) l TCP33

Advanced Bayesian Modeling

Organizer: Junhao Pan (Sun Yat-sen University)

Chair: Junhao Pan (Sun Yat-sen University)

Bayesian Analysis of the Functional-coefficient Autoregressive Heteroscedastic Model

Jingheng Cai

Sun Yat-sen University, China. *caijheng@mail.sysu.edu.cn*

Abstract: In this paper, we propose a new model called the functional-coefficient autoregressive heteroscedastic (FARCH) model for nonlinear time series. The FARCH model extends the existing functional-coefficient autoregressive models and double-threshold autoregressive heteroscedastic models by providing a flexible framework for the detection of nonlinear features for both the conditional mean and conditional variance. We propose a Bayesian approach, along with the Bayesian P-splines technique and Markov chain Monte Carlo algorithm, to estimate the functional coefficients and unknown parameters of the model. We also conduct model comparison via the Bayes factor. The performance of the proposed methodology is evaluated via a simulation study. A real data set derived from the daily S&P 500 Composite Index is used to illustrate the methodology.

Bayesian Adaptive Lasso for Multivariate Generalized Linear Model with Latent Variables

Xiang-Nan Feng

The Chinese University of Hong Kong, Hong Kong. *fengxiangnan123@gmail.com*

Abstract: We consider a multivariate generalized linear model (GLM) with latent variables to investigate the effects of observable and latent explanatory variables on the mixed type responses of interest. Various types of responses, such as continuous, count, ordinal, and multinomial variables are considered in the regression. A generalized confirmatory factor analysis model that is able to handle mixed type correlated observed indicators is jointly analyzed with the GLM to characterize the latent variables. To cope with the complicated model structure, a Bayesian adaptive lasso procedure is developed for simultaneous model estimation and selection. We evaluate the performance of the proposed method through extensive simulation studies and a real data example.

Bayesian Approaches in Analyzing Earthquake Catastrophic

Xuejun Jiang

South University of Science and Technology of China, China. *jiangxj@sustc.edu.cn*

Abstract: Extreme value theory is widely used to analyze the catastrophic risk. According to the theory, the limiting distribution of the excess value over a certain threshold is the generalized Pareto distribution (GPD). Based on the extreme value theory, we can estimate the tail behaviors of GPD. However, the central behavior is also important as it may affect estimation of model parameters in GPD, and the evaluation of catastrophic insurance premiums also depends on the central behavior. In this paper, we propose four mixture models to model the earthquake catastrophic loss and develop Bayesian approaches to estimate the unknown parameters and the threshold in these mixture models. In particular, we use MCMC methods to calculate Bayesian estimates of model parameters and obtain Deviance information criterion (DIC) values for model comparisons. We also analyze the earthquake losses of Yunnan province to illustrate the proposed methods. Our results indicate that estimation of the threshold as well as the shape and scale of GPD are quite different. Under different confidence levels, we calculate Value-at-Risk (VaR) and expected shortfall (ES) for the proposed mixture models to make decisions for risk management analysis.

Bayesian Transformation Quantile Regression

Pengfei Liu

Jiangsu Normal University, China. *liupengfei@jsnu.edu.cn*

Abstract: As useful techniques, transformation models have attracted significant attention from various fields. We develop a Bayesian approach for analyzing transformation quantile regression. We also consider types of missingness, such as missing completely at random, missing at random, and missing not at random, are discussed under the transformation framework. The empirical performance of our methodology is examined via simulation studies.

Efficient Transformations for Exploring MCMC Sampler on a Family of Banana-shaped Distributions

Maolin Pan

Nanjing University, China. *mlpan@nju.edu.cn*

Abstract: In this study, a family of banana-shaped distribution as a model for general banana-shaped distributions and a corresponding transformation-based MCMC algorithm will be proposed. The new method is better than traditional methods in terms of mixing and convergence rate for banana-shaped distributions. Simulations and examples of the new method are given and some corresponding comparisons are made with existing methods.

Thu, June 30 (10:30-12:10) | TCP07

Recent Advances in High Dimensional Estimation Theory

Organizer: Gourab Mukherjee (University of Southern California)

Chair: Gourab Mukherjee (University of Southern California)

Bayes Projection and its Applications to High-dimensional Problems

Fumiyasu Komaki

The University of Tokyo, Japan. *komaki@mist.i.u-tokyo.ac.jp*

Abstract: Constructing methods for predictive densities based on parametric models are considered. The Bayes projection of a predictive density is defined as a Bayesian predictive density based on a prior minimizing a divergence from it to the original predictive density. The Bayes projection is superior to the original predictive density under some regularity conditions. Several numerical algorithms to construct Bayes projection and their applications are discussed.

A Nonparametric Bayesian Approach for Sparse Sequence Estimation

Yunbo Ouyang

University of Illinois at Urbana-Champaign, United States. *youyang4@illinois.edu*

Abstract: A nonparametric Bayes approach is proposed for the problem of estimating a sparse sequence based on Gaussian random variables. We adopt the popular two-group prior with one component being a point mass at zero, and the other component being a Gaussian distribution. Although the Gaussian prior has shown to be suboptimal, we find that with an adaptive choice on the Gaussian mean, we can show that the posterior distribution has the desirable asymptotic behavior, e.g. it concentrates on balls with the desired minimax rate.

Optimal Community Detection in Degree-corrected Stochastic Block Models

Anderson Ye Zhang

Yale University, United States. *ye.zhang@yale.edu*

Abstract: In this talk, we first examine the minimax rates for community detection in degree-corrected stochastic block models (DCBM). Guided by the minimax rates, we then discuss a computationally tractable algorithm for achieving them adaptively over a wide range of parameter spaces for DCBM.

Group-linear Empirical Bayes Estimation of a Heteroscedastic Normal Mean

Asaf Weinstein

Stanford University, United States. *asafw.at.stanford@gmail.com*

Abstract: We revisit a classic problem: X_i ~N(θ_i,V_i) indep, V_i known, i=1,...,n, and the goal is to estimate the (nonrandom) means θ_i under sum of squared errors. When the variances are all equal, linear empirical Bayes estimators which model the true means as i.i.d. random variables lead to (essentially) the James-Stein estimator, and have strong frequentist justifications. In the heteroscedastic case such empirical Bayes estimators are less adequate if the V_i and θ_i are empirically dependent. We suggest a new empirical Bayes procedure which groups together observations with similar variances and applies a spherically symmetric estimator to each group separately. Our estimator is exactly minimax and at the same time asymptotically achieves the risk of a stronger oracle than the usual one. The motivation for the new estimator comes from extending a compound decision theory argument from equal variances to unequal variances.

This is joint work with Larry Brown, Zhuang Ma and Cun-Hui Zhang.

Thu, June 30 (10:30-12:10) I TCP05

High-Dimensional Data Analyses with Application in Biomedical Studies

Organizer: Yi Li (University of Michigan)

Chair: Ming Gao Gu (The Chinese University of Hong Kong)

Covariance-insured Screening Methods for Ultrahigh Dimensional Variable Selection

Kevin (Zhi) He

University of Michigan, United States. *kevinhe@umich.edu*

Abstract: Effective screening methods are crucial to the analysis of big biomedical data. The popular sure independence screening relies on restricted assumptions such as the partial faithfulness condition, e.g., the partial correlation between outcome and covariates can be inferred from their marginal correlation. However, such a restrictive assumption is often violated, as the marginal effects of predictors may be quite different from their joint effects, especially when the covariates are correlated. We propose a covariance-insured screening (CIS) framework that utilizes the dependence among covariates and identify important features that are likely to be missed by marginal screening procedures such as sure independence screening. The proposed framework encompasses linear regression models, generalized linear regression models and survival models.

Detecting Association to Precision Networks via Conditional Multi-type Graphical Models

Yanming Li

University of Michigan, United States. *liyanmin@umich.edu*

Abstract: Understanding how genes regulate each other and how their regulations are associated with genomic markers with respect to individual patients can help uncover the mechanism of underlying biological or disease process at DNA level. Conditional graphical models are commonly used in simultaneously learning the gene regulatory network and recovering the association signals. Most of current conditional graphical models assume a homogeneous response network structure and only model the responses conditional means on the predictors. Also the current methods are not able to handle multi-type responses networks, therefore limit their applications in modern biomedical studies. We propose a multi-type conditional graphical model which allows heterogeneous patient level responses networks, mixture of types of responses and can accurately and effectively recover the associations of responses networks to high-dimensional biomarkers. The proposed method is computationally inexpensive and its finite sample properties are investigated both theoretically and empirically. We apply the method to the TCGA data to better understand the cancer-triggering biological pathways at molecular level.

Detecting Rare and Faint Signals via Thresholding Maximum Likelihood Estimators

Yumou Qiu

University of Nebraska–Lincoln, United States. *yumouqiu@unl.edu*

Abstract: Motivated by the analysis of RNA sequencing (RNA-seq) data for genes differentially expressed across multiple conditions, we consider detecting rare and faint signals in high-dimensional response variables. We address the signal detection problem under a general framework, which includes generalized linear models for count-valued responses as special cases. We propose a test statistic that carries out a multi-level thresholding on maximum likelihood estimators (MLEs) of the signals, based on a new Cramer type moderate deviation result for multi-dimensional MLEs. Based on the multi-level thresholding test, a multiple testing procedure is proposed for signal identification. Numerical simulations and a case study on maize RNA-seq data are conducted to demonstrate the effectiveness of the proposed approaches on signal detection and identification.

Automatic Detection of Significant Areas for Functional Data with Directional Error Control

Peirong Xu

Southeast University, China. *xupeirong@seu.edu.cn*

Abstract: To detect differences between mean curves of two samples in longitudinal study or functional data analysis, we usually need to partition the temporal or spatial domain into several pre-determined sub-areas. In this paper we apply the idea of large-scale multiple testing to find the significant sub-areas automatically in a general functional data analysis framework. A nonparametric Gaussian process regression model is introduced for two-sided multiple tests. We derive an optimal test which controls directional false discovery rates and propose a procedure by approximating it on a continuum. The proposed procedure controls directional false discovery rates at any specified level asymptotically. In addition, it is computationally inexpensive and able to accommodate different time points for observations across the samples. Simulation studies are presented to demonstrate its finite sample performance. We also apply it to an executive function research in children with Hemiplegic Cerebral Palsy and extend it to the equivalence tests.

Set-based Test for Gene-environment Interaction

Baqun Zhang

Renmin University of China, China. *zhangbaqun@ruc.edu.cn*

Abstract: There has been tremendous emphasis on searching for interactions between genetic factors and environmental exposures. Gene-environment interactions (GxE) are typically based on testing the interaction between each single-nucleotide polymorphisms (SNP) and an environmental variable separately, with adjustment for multiple testing. However, the interaction process is probably far more complex than looking for "single locus vs. environment factor" analysis. We propose a novel statistical approach to test for gene-environment interaction between an environmental factor and a high dimensional set of genetic variants for longitudinal studies, with the consideration of potential time dependency and correlation in the outcomes measured on the same subject. The method integrates the entire genotype-environment-phenotype information contained in a longitudinal study through a region based test. Nonparametric modeling of the environmental exposure is incorporated to alleviate the problem of misspecification of the main or interaction effect, leading to more robust type-I error rate and superior power. As the number of SNPs in a target region can be very large, dimension reduction method is further proposed, which selects and adjusts for the main effect of genetic variants to achieve numerical feasibility, controlled type I error probability and improvement in power. The performance of the method will be evaluated through simulation studies and illustrated by real data analysis.

Thu, June 30 (10:30-12:10) l CP11

Big Data/Network Session 1

Chair: Xuexia Wang (University of Wisconsin-Milwaukee)

Monitoring the Results of Cardiac Surgery Based on 3 or More Outcomes by Variable Life-adjusted Display

Fah Fatt Gan

National University of Singapore, Singapore. *staganff@nus.edu.sg*

Abstract: The variable life-adjusted display (VLAD) was proposed as a simple graphical display for monitoring results of cardiac surgery, adjusted by preoperative risks of patients. The VLAD is now popularly used all over the world. It 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. This results in a great loss of information and hence a VLAD that does not reveal the true results. We develop a VLAD that is based on three or more outcomes and this refined procedure will reveal more accurately and fairly the results. This VLAD will be referred to as the new VLAD.

An Adaptive 3-D Image Denoising Framework

Partha Sarathi Mukherjee

Boise State University, United States. *parthamukherjee@boisestate.edu*

Abstract: Recent increase in the use of 3-D magnetic resonance images (MRI) and analysis of functional magnetic resonance images (fMRI) makes 3-D imaging very important. Observed 3-D images often contain noise which should be removed in such a way that important image features, e.g., edges, edge structures, and other image details should be preserved, so that subsequent image analyses are reliable. Most image denoising methods in the literature are for 2-D images. However, their direct generalization to 3-D images cannot preserve complicated edge structures well, because, the edge structures in a 3-D edge surface can be much more complicated than the edge structures in a 2-D edge curve. Moreover, the amount of smoothing should be determined locally, depending on local image features and local signal to noise ratio, which is much more challenging in 3-D images due to large number of voxels. This paper proposes an efficient 3-D image denoising procedure based on local clustering of the voxels. This method provides a framework for determining the size of bandwidth and the amount of smoothing locally by data driven procedures. Numerical studies show that it works well in many real world applications.

New Types of Shrinkage Estimators of Poisson Means

Genso(Yuan-Tsung) Watanabe(Chang)

Mejiro University, Japan. *cho@mejiro.ac.jp*

Abstract: In estimating $p(\geq 2)$ independent Poisson means, Clevenson and Zidek (1975) have suggested a class of estimators that shrink the unbiased estimators to the origin and have shown that the shrinkage estimators dominate the unbiased ones under the normalized squared error loss.

In this article we suggest new types of estimators of p independent Poisson means, which shrink the unbiased estimators to their minimum, or more generally to some order statistics. We use the partial summation method to show the dominance results. We also discuss the two-way contingency table saturated Poisson model, for each row shrinks to their order statistics to invalid the interaction effects.

Further, we apply the suggested method to the multiplicative Poisson models and suggest some types of shrinkage estimators which shrinks to the order statistics.

Spatial Autoregressive Model Estimation for Large-scale Social Networks

Yingying Ma

Beihang University, China. *mayingying_11@163.com*

Abstract: Due to the rapid development of Social Networks, the usefulness of the spatial autoregressive model has been recognized and popularly applid to explore social network structures. However, traditional estimation methods are practically infeasible if network size is huge (e.g., Facebook, Twitter, Sina Weibo, WeChat, QQ, etc). We propose here a novel estimation approach, which has reduce the computational complexity from n^3 to n. We theoretically proved that the proposed estimation method is consistency. In addition to that, the proposed method can be readily applied to sampled network data. Numerical studies based on both simulated and real datasets are presented.

Finite Sample Covariance Estimation Based on Graphical Model

Chaojie Wang

The Chinese University of Hong Kong, Hong Kong. *wang910930@163.com*

Abstract: Covariance estimation in high-dimension ($p>n$) case draw much attention for the past few years. Related problems are found in the fields of probability, statistics and machine learning, etc. Traditional thresholding methods would lose positive-definiteness easily, which is a necessary property for covariance matrix. Here, we propose a sparsity-controlled estimate, which ensures positive-definiteness, under MLE criterion in finite sample and also an efficient algorithm to attain it. Graph model provides a visible explanation for relationship between Gaussian variables and a powerful tool to deal with our problem.

Thu, June 30 (13:30-15:10) l DL04 l Sponsor: IMS

From Cells to Populations: Modeling and Inference for Genomic Data

Chair: Susan Wilson (The University of New South Wales and The Australian National University)

Distinguished Lecturer: Elizabeth Thompson (University of Washington)

Modelling and Inference of Co-ancestry in Populations

Elizabeth Thompson

University of Washington, United States. *eathomp@u.washington.edu*

Abstract: There are many HMM approaches to inference of co-ancestry or identity-by-descent (IBD) between pairs of individuals under simple population models. However, these models are limited in two major ways. First, pairwise IBD may not correspond to any valid IBD jointly among multiple individuals. Second, models typically ignore dependence of allelic types (LD) across a chromosome. A new model for local population haplotype frequencies is presented, and it is shown how these may be used to reweight IBD realizations to adjust for LD.

Three Dimensional Chromatin Structure and Spatial Gene Regulation

Shili Li

The Ohio State University, United States. *shili@stat.ohio-state.edu*

Abstract: The expression of a gene is usually controlled by the regulatory elements in its promoter region. However, it has long been hypothesized that, in complex genomes, such as the human genome, a gene may be controlled by distant enhancers and repressors. A recent high throughput molecular technique, Hi-C, that uses formaldehyde cross-linking coupled with massively parallel sequencing, enables detections of genome-wide physical contacts between distant loci. Such communication is achieved through spatial organization (looping) of chromosomes to bring genes and their regulatory elements into close proximity. The availability of such data makes it possible to reconstruct the underlying three-dimensional (3D) spatial chromatin structure and to study spatial gene regulation. In this talk, I will describe a number of statistical methods, including truncated Random effect EXpression (tREX) method for inference on the locations of genomic loci in a 3D Euclidean space. Results from Hi-C data will be visualized to illustrate spatial regulation and proximity of genomic loci that are far apart in their linear chromosomal locations.

Advances of Bayesian Nonparametrics in Population Genetics of Infectious Diseases

Vladimir Minin

University of Washington, United States. *vminin@uw.edu*

Abstract: Estimating evolutionary trees, called phylogenies or genealogies, is a fundamental task in modern biology. Once phylogenetic reconstruction is accomplished, scientists are faced with a challenging problem of interpreting phylogenetic trees. In certain situations, a coalescent process, a stochastic model that randomly generates evolutionary trees, comes to rescue by probabilistically connecting phylogenetic reconstruction with the demographic history of the population under study. An important application of the coalescent is phylodynamics, an area that aims at reconstructing past population dynamics from genomic data. Phylodynamic methods have been especially successful in analyses of genetic sequences from viruses circulating in human populations. From a Bayesian hierarchal modeling perspective, the coalescent process can be viewed as a prior for evolutionary trees, parameterized in terms of unknown demographic parameters, such as the population size trajectory. I will review Bayesian nonparametric techniques that can accomplish phylodynamic reconstruction, with a particular attention to analysis of genetic data sampled serially through time, because such data are common in infectious disease epidemiology. I will illustrate Bayesian nonparametric phylodynamic reconstruction methods using genetic data from seasonal human influenza.

Thu, June 30 (13:30-15:10) l DL15 l Sponsor: India

Random Networks

Chair: Arup Bose (Indian Statistical Institute)

Distinguished Lecturer: Rahul Roy (Indian Statistical Institute)

Drainage Networks and the Brownian Web

Rahul Roy

Indian Statistical Institute, India. *rahul.isid@gmail.com*

Abstract: River basin geomorphology is a very old subject of study initiated by Horton (1945). Various statistical models of drainage networks have been proposed. Each such model is a random directed graph with its own nuances. In recent years physicists have been interested in these models because of the commonality of such branching networks in many areas of statistical physics (see Rodrguez-Iturbe and Rinaldo (1997) for a detailed survey). We discuss the geometric features of one such model and also its scaling limit. The scaling limit of this model is the Brownian web, which has lately been the focus of extensive study among probabilists. Using this scaling limit we formalize Hack's law, a proportionality relation between the length and the watershed area of a river widely used in geology.

This is joint work with Kumarjit Saha and Anish Sarkar.

Rumor Spreading in Dynamic Networks

Marco Isopi

Sapienza University of Rome, Italy. *isopi@mat.uniroma1.it*

Abstract: Randomized gossip is one of the most popular way of disseminating information in large scale networks.

In the so-called Push protocol, every informed node selects, at every time step, a neighboring node uniformly at random and forwards the information to this node. This protocol is known to complete information spreading in O(log n) time steps with high probability in several families of n-node static networks.

I will discuss the Push protocol in dynamic networks. We will consider a discrete time Markov evolution where each non-edge appears with some probability p(n), while an existing edge dies with probability q(n). We will show that for several regimes, i.e. asymptotic behaviors of p(n) and q(n), the Push protocol completes information spreading in O(log n) time steps with high probability. Interestingly this holds even when the network is, w.h.p., not connected at every time step.

We will also review regimes where completion in O(log n) steps does not happen and discuss what can be said when information may be forgotten by a node.

Application of Random Graphs in Epidemiology and Economics

Farkhondeh Alsadat Sajadi

University of Isfahan, Iran. *farkhondeh_sajadi@yahoo.com*

Abstract: In this talk, we discuss some applications of random graph models in two branches of science, namely, epidemiology and economics. We study one of the problems of epidemics spreading on contact networks which is finding the number of individuals that eventually become infected (and removed). We consider a simple SIR infection spread model on a finite network of agents, and our goal is to determine an approximation of the total amount of infection without specifying the underlying network. Starting with a fixed set of initial infected vertices the infection spreads in discrete time steps, where each infected vertex tries to infect its neighbors with a fixed probability $\beta \in (0, 1)$, independently of others. It is assumed that each infected vertex dies out after an unit time and the process continues till all infected vertices die out. We find a simple lower bound on the expected number of ever infected vertices using breath-first search algorithm.

In context of economics, we discuss about the trading system as a scale-free network. Every product is traded in a network. Countries and trade volume can be interpreted as the nodes and weighted directed edges of these networks. We study the reason why scale-free small-world network is a good model for describing the trading network and why the distribution of trading links should follow a power-law.

Thu, June 30 (13:30-15:10) l IP48

Recent Advances in Lifetime Data Analysis

Organizer: Mei-Ling Ting Lee (University of Maryland)

Chair: Xingqiu Zhao (The Hong Kong Polytechnic University)

Causal Inference and Time

Odd O. Aalen

University of Oslo, Norway. *o.o.aalen@medisin.uio.no*

Abstract: Why is causality important? In medicine the obvious reason for this is that medical personnel constantly make important decisions. Medical doctors, in particular, intervene in people’s lives, and many of these interventions have their basis in statistical studies. But the concept of intervention is not a part of classical statistics, in fact, most classical statistics used to be silent on intervention. It was first Judea Pearl and coauthors who pointed out that we need a mathematical language to deal with interventions and that causality is mainly about intervention.

Today, causal inference is everywhere. A weakness of the way the field presents itself today is the general absence of time. Behind the causal effects there are some mechanisms operating, and, obviously, these operate in continuous time. We need to understand the implications of this. This is not least the case in the modern plethora of mediation studies. Clearly, mediation must be seen as a dynamic concept.

The issue of time is particularly relevant in survival and event history analysis. The modern development of causal inference has strong implications for this area. E.g., the implicit conditioning on survival which is present in the definition of intensity processes and hazard rates opens for collider effects. We shall look at the implications of this.

Survival and Quality of Life

Catherine Louise Huber-Carol

Université Paris Descartes, France. *catherine.huber@parisdescartes.fr*

Abstract: Survival data analysis and reliability have many common points and the concepts that are studied are comparable. But the data have often different properties as we are not in the same experimental field. For example, we cannot do "accelerated" experiments on patients. Also the patients are not only "objects" of study but also "subjects". And while we try to compare the effects of two drugs to cure a particular disease, the patient may decide at any time to stop his participation in the experiment, which leads to what we call "right censoring" of the data which are only partially observed. If we add to this feature the possibility of "truncation" (only a part of the sample is observed, some subjects are not included in the sample) and the fact that the size of the patient sample is often limited to small numbers, we can see that there are multiple differences in the available data in survival analysis as compared to reliability analysis. However, the same models can be used. For example, even if we cannot "accelerate" experiments on patients by enforcing the "stresses" applied to them, we can use such an accelerated model in order to understand whether a specific exposure to pollution is accelerating the onset of a specific disease. We give here an example where the occupational exposure to asbestos is shown to accelerate the onset of lung cancer. The aim of this talk is to show how one can deal with the censoring and truncation inconveniences, in both case of only one terminal event(1,3,4) and also when several different outcomes may happen(5,6). But, as the aim of survival analysis is not only to try to increase the length of the pure survival time (to death) but also the survival time "free of disease", this leads to estimate the number of "years free of disease" lost(2,7) due to environmental or behavioural factors for the sake of prevention.

Nonparametric Analysis of the Dependence Structure for Recurrent Gap Time Data

Shu-Hui Chang

National Taiwan University, Taiwan. *shuhui@ntu.edu.tw*

Abstract: In this study, we utilize the ordinal nature of recurrent gap times to introduce serial association measures to identify and evaluate possible dependence structures for recurrent event processes. Nonparametric estimators of the serial association measures are developed by generalized the existing estimators of rank-based pairwise association measures in which inverse probability of censoring weights is used to handle the induced dependent censoring. We further develop nonparametric tests for detecting and comparing the dependence patterns. The asymptotic multivariate normal distributions and the estimated asymptotic variance-covariance matrices of the proposed methods are derived from U-statistics theories. A simulation study is presented to assess the finite-sample properties of the proposed methods. An application of schizophrenia data suggests that the dependence patterns among recurrent gap times between relapses of the illness for patients vary with age at the onset of schizophrenia.

Estimating Model-based Attributable Risk Functions in the Asian Cohort Consortium

Ying Qing Chen

Fred Hutchinson Cancer Research Center, United States. *yqchen@fhcrc.org*

Abstract: In disease prevention research, population attributable fraction (PAF) has been a useful measure to quantify the additional disease risk in a population associated with certain modifiable risk factors. When the disease outcome is of the time-to-event type, the PAF can naturally be extended as a function of time, or the so-called attributable risk function (ARF). In this paper, we develop model-based estimating and inference procedures for the ARF under the additive hazards model, which has been an important alternative to the widely used Cox proportional hazards model but carries its own public health implication. Our methods will be applied to the data assembled by the Asian Cohort Consortium, to estimate the time-varying public health impact of smoking on the mortality risk.

Thu, June 30 (13:30-15:10) l IP46

Recent Developments in Survival Analysis and Personalized Medicine

Organizer: Wenbin Lu (North Carolina State University)

Chair: Rui Song (North Carolina State University)

Concordance-assisted Learning for Estimating Optimal Individualized Treatment Regimes

Wenbin Lu

North Carolina State University, United States. *lu@stat.ncsu.edu*

Abstract: A new concordance-assisted learning (CAL) is presented for estimating optimal individualized treatment regimes. First, we will introduce a type of concordance function for prescribing treatment and propose a robust rank regression method for estimating the concordance function. Then, we will discuss the proposed CAL methods for estimating optimal treatment regimes that maximize the concordance function, named prescriptive index, and for searching the optimal threshold. Moreover, we will discuss the convergence rates and asymptotic distributions of the proposed estimators for parameters in the prescriptive index and the optimal threshold. Finally, we will present some simulations and an application to an AIDS data to illustrate the practical use and effectiveness of the proposed methodology.

Semiparametric Structural Equation Models with Latent Variables for Right-censored Data

Kin Yau Wong

The University of North Carolina at Chapel Hill, United States. *alexwky@live.unc.edu*

Abstract: Structural equation modeling is commonly used to capture complex structures of relationships among multiple variables, both latent and observed. In this presentation, we discuss a general class of structural equation models with a semiparametric component for potentially censored survival times. We consider nonparametric maximum likelihood estimation and devise a combined Expectation-Maximization and Newton-Raphson algorithm for its computation. We establish conditions for model identifiability and prove the consistency, asymptotic normality, and semiparametric efficiency of the estimators. Finally, we demonstrate the satisfactory performance of the proposed methods through simulation studies and provide application to a motivating cancer study that contains a variety of genomic variables.

Flexible Modeling of Bivariate Recurrent Events Data

Limin Peng

Emory University, United States. *lpeng@sph.emory.edu*

Abstract: Recurrent events are frequently observed in biomedical studies, and often there are more than one type of events of interests. Marginal analysis of each type of recurrent event is useful but cannot address questions on the relationship between different types of recurrent events. In this work, we study a dynamic association model built upon the stochastic processes embedded with bivariate recurrent events data. The proposed estimation can be implemented by an efficient and stable algorithm. We investigate the asymptotic properties of the proposed estimator, and develop proper inference procedures. Our proposals are illustrated via simulation studies and an application to a registry dataset.

A Semiparametrically Efficient Estimator of the Time-varying Effects for Survival Data with Time-dependent Treatment*

Huazhen Lin

Southwestern University of Finance and Economics, China. *linhz@swufe.edu.cn*

Abstract: The timing of time-dependent treatment---e.g., when to perform kidney transplantation---is an important factor for evaluating treatment efficacy. A naive comparison between the treatment and nontreatment groups, while ignoring the timing of treatment, typically yields results that might biasedly favor the treatment group, as only patients who survive long enough will get treated. On the other hand, studying the effect of time-dependent treatment is often complex, as it involves modeling treatment history and accounting for the possible time-varying nature of the treatment effect. We propose a varying-coefficient Cox model that investigates the efficacy of time-dependent treatment by utilizing a global partial likelihood, which renders appealing statistical properties, including consistency, asymptotic normality and semiparametric efficiency. Extensive simulations verify the finite sample performance, and we apply the proposed method to study the efficacy of kidney transplantation for end-stage renal disease patients in the U.S. Scientific Registry of Transplant Recipients (SRTR).

*Joint work with Zhe Fei, Yi Li

Thu, June 30 (13:30-15:10) | IP53

Some Recent Developments in High-Frequency Financial Econometrics

Organizer: Bingyi Jing (The Hong Kong University of Science and Technology)

Chair: Bingyi Jing (The Hong Kong University of Science and Technology)

Testing the Equality of Large U-statistic Based Correlation Matrices

Xinsheng Zhang

Fudan University, China. *xszhang@fudan.edu.cn*

Abstract: In this talk, we provide a framework for testing the equality of two large U-statistic based correlation matrices, which include the rank-based correlation matrices as special cases. Our approach exploits extreme value statistics and the Jackknife estimator for uncertainty assessment and is valid under a fully nonparametric model. Theoretically, we develop a theory for testing the equality of U-statistic based correlation matrices. We then apply this theory to study the problem of testing large Kendall's tau correlation matrices and demonstrate its optimality. For proving this optimality, a novel construction of least favourable distributions is developed for the correlation matrix comparison. This is joint work with Cheng Zhou, Fang Han, and Han Liu.

On the Integrated Systematic and Idiosyncratic Volatility with Large Panel High-frequency Data

Xinbing Kong

Soochow University, China. *kongxinbing@suda.edu.cn*

Abstract: In this paper, we separate the integrated volatility of an individual Ito process into the integrated systematic and idiosyncratic volatility, and estimate them by aggregation of local factor analysis with large dimensional high-frequency data. It is shown that, when both the sampling frequency n and the dimensionality p go to infinity, our estimators of the integrated systematic and idiosyncratic volatility could be $\sqrt{n}$ consistent, the best rate achieved in estimating the integrated volatility readily identified even with univariate high-frequency data. We also present an estimator of the integrated idiosyncratic volatility matrix under some sparsity assumption which typically does not hold for the integrated volatility matrix (the sum of the integrated systematic and idiosyncratic volatility matrices). It is proved that the estimator converges in the operator norm at the rate of $s_0(p)(\frac{1}{\sqrt{p}}+\frac{\sqrt{\log p}}{\sqrt{n}})^{1-q}$ where $s_0(p)$ is a measure of sparsity. Related to the portfolio selection theory, we also present a factor-based estimator of the inverse of the integrated volatility matrix which converges to the realized population counterpart with the rate of $s_0(p)(\frac{1}{\sqrt{p}}+\frac{\sqrt{\log p}}{\sqrt{n}})^{1-q}+\frac{(\log p)^{1/4}}{n^{3/8-\varepsilon}}+\frac{1}{n^{1/4}p^{3/8}}$. Numerical studies including the Monte-Carlo experiments and real data analysis justify the performance of our estimators.

Testing for Presence of Leverage Effect Under High Frequency

Zhi Liu

University of Macau, Macao. *liuzhi@umac.mo*

Abstract: In this paper, we propose a test for deciding whether the correlation of a discretely-observed semi-martingale and its quadratic variation (refers to the leverage effect in the financial econometric) equals to zero. The asymptotic setting is based on observations within a long time interval with mesh of the observation grid shrinking to zero. The test is based on forming a sequence of studentized statistics whose distributions are asymptotically normal locally over blocks of shrinking time span, and the collecting the sequence based on the whole data set. The asymptotic behaviour of the local studentized statistics is obtained from a similar result in a global setting using the third power variation of the underlying process. We derive the asymptotic distribution of the proposed test statistic under the null hypothesis of zero leverage effect and show that the test has asymptotic power of one against fixed alternatives of processes with non-zero leverage effect. Finally, simulation study verifies the finite sample performance of the test.

Adaptive Thresholding for Large Volatility Matrix Estimation Based on High-frequency Financial Data

Cuixia Li

Lanzhou University, China. *licuixia@lzu.edu.cn*

Abstract: Universal thresholding estimators have been developed to estimate the large sparse integrated volatility matrix. Since the integrated volatility matrix often has entries with a wide range of variability, universal thresholding estimators do not take the varying entries into consideration and have unsatisfactory performances. This paper investigates adaptive thresholding estimation of large volatility matrix. We first construct an estimator for the asymptotic variance of the pre-averaging realized volatility estimator to develop an adaptive thresholding estimator of the large volatility matrix. It is shown that the adaptive thresholding estimator can achieve the optimal rate of convergence over the class of the sparse integrated volatility matrix when both the number of assets and sample size are allowed to go to infinity, while the universal thresholding estimator can achieve only the sub-optimal convergence rate. The simulation study is conducted to check the finite sample performance of the adaptive thresholding estimator.

Thu, June 30 (13:30-15:10) l IP33 l Sponsor: India

Analysis of Spatial and Spatio-Temporal Data

Organizer: Debasis Sengupta (Indian Statistical Institute)

Chair: Marc G. Genton (King Abdullah University of Science and Technology)

Modeling Tangential Vector Fields on the Sphere

Debashis Paul

University of California, Davis, United States. *debashis@wald.ucdavis.edu*

Abstract: Random vector fields on the sphere appear naturally in many geophysical processes. Vector processes that are observed at the ground level and are tangential to the earth's surface include the surface wind velocity field, the horizontal component of the earth's magnetic field, etc. Obtaining a modeling framework that respects the curvature of the spherical domain, while having physical interpretations, is the main focus of this talk. We introduce a new class of stochastic processes, named Tangent Matérn Model (TMM), based on a Helmholtz-Hodge decomposition of vector fields that achieves both these goals. We also propose a method for estimating the TMM based on the likelihood framework. The proposed method is applied to data on the wind velocity field at sea surface collected from a satellite-based scatterometry survey. The talk is based on joint work with Minjie Fan (UC Davis), Thomas Lee (UC Davis) and Tomoko Matsuo (NOAA).

A Novel Nonparametric Threshold-free Method to Produce Functional MRI Activation Maps

Rajesh Ranjan Nandy

UNT Health Science Center, United States. *Rajesh.Nandy@unthsc.edu*

Abstract: A primary objective of spatio-temporal fMRI data analysis is to identify the active voxels in the brain while a task is being performed. A usual method to achieve this objective is to use a test statistic which is a function of the observed signal and declare a voxel as active depending on whether the observed value of the test statistic meets or exceeds a pre-determined threshold usually based on a family-wise error rate (FWER) or false discovery rate (FDR). This approach is well-established but suffers from a big weakness: a subjective choice of the pre-determined threshold to produce the activation map. Clearly the activation maps will look very different for varying choices of thresholds where larger activation blobs can be achieved by sacrificing specificity. In the proposed method, we no longer need to choose a threshold since it is optimally chosen from the data so that total misclassification rate of voxels classified as active or inactive is minimized.

A Statistical Description of the Spatial Extent of a Spell of Rainfall

Subrata Kundu

The George Washington University, United States. *kundu@gwu.edu*

Abstract: The spatial extent of a spell of rainfall is a connected region with positive rainfall at a particular time. The probabilistic behavior of the spatial extent of a spell of rainfall and various attributes of it, are issues of interest in meteorological studies. While the spatial extent can be viewed as a shape object, scale and rotational invariance of the shape are not necessarily desirable attributes from meteorological considerations. For modeling objects of the above type, we propose a computationally efficient multivalued functional representation of the shape of the rainfall region and an appropriate linear space, with an associated distance measure. While a probability density function does not exist in this situation, it is possible to develop a meaningful surrogate for a density when functional data are considered in the space determined by eigenfunctions in a principal component analysis. We develop a method for deriving the probability distribution of a general functional of the shape from the surrogate probability density function for the shape and propose a nonparametric method to estimate this probability distribution. Strong consistency of the proposed estimator is established. This method is used to analyze an open access satellite data set over the West Bengal, India.

Keywords and phrases: Functional principal component analysis, Shape analysis, Kernel density estimation.

Zero Expectile Processes and Bayesian Spatial Regression

Anandamayee Majumdar

Soochow University, China. *anandamayee.majumdar@gmail.com*

Abstract: We introduce new classes of stationary spatial processes with asymmetric, sub-Gaussian marginal distributions using the idea of expectiles. We derive theoretical properties of the proposed processes. Moreover, we use the proposed spatial processes to formulate a spatial regression model for point-referenced data where the spatially correlated errors have skewed marginal distribution. We introduce a Bayesian computational procedure for model fitting and inference for this class of spatial regression models. We compare the performance of the proposed method with the traditional Gaussian process-based spatial regression through simulation studies and by applying it to a data set on air pollution in California.

Thu, June 30 (13:30-15:10) | CP12

Big Data/Network Session 2

Chair: Fah Fatt Gan (National University of Singapore)

Enhanced Construction of Gene Regulatory Networks using Hub Gene Information

Donghyeon Yu

Keimyung University, South Korea. *dyu3@kmu.ac.kr*

Abstract: Gene regulatory networks reveal how genes work together to carry out their biological functions. Reconstructions of gene networks from gene expression data greatly facilitate our understanding of underlying biological mechanisms and provide new opportunities for biomarker and drug discoveries. In gene networks, a gene that has many interactions with other genes is called a hub gene, which usually plays an essential role in gene regulation and biological processes. In this study, we developed a method for reconstructing gene networks using a partial correlation-based approach that incorporates prior information about hub genes. Through simulation studies and two real-data examples, we compare the performance in estimating the network structures between the existing method and the proposed method. In simulation studies, we show that the proposed procedure reduces errors in estimating network structures compared to the sparse partial correlation estimation (SPACE) method. When applied to Escherichia coli, the regulation network constructed by our proposed approach is more consistent with current biological knowledge than the SPACE method. Furthermore, application of the proposed method in lung cancer has identified hub genes whose mRNA expression predicts cancer progress and patient response to treatment. We have considered incorporating hub gene information in estimating network structures, which can improve the performance of the existing methods.

A Popularity Scaled Latent Space Model for Network Structure Formulation

Xiangyu Chang

School of Management, Xi'an Jiaotong University, China. *xiangyuchang@gmail.com*

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

Network Dynamics Detection using Liquid Association

Tianwei Yu

Emory University, United States. *tianwei.yu@emory.edu*

Abstract: Detecting complex relations in expression data may reveal important regulatory mechanisms in the biological system. However, the study of relations between three or more genes may involve too big a search space, i.e. too many possible combinations, and could generate spurious results. Inspired by the recent trend of analyzing gene expression data based on the structure of existing biological networks, we propose a new method to study gene expression dynamics. The method is named LANDD: Liquid Association for Network Dynamics Detection. It uses Liquid Association (LA) to detect three-way interactions, while constraining the analysis to regions of the existing biological network using the ego-network concept. While reducing the search space of gene triplets substantially, the method produces easily interpretable results because of its focus on sub-networks of functionally related genes. Using a real gene expression dataset and the human protein-protein interaction network, we demonstrate the method links network regions of distinct biological processes together with new and plausible functional implications.

Novel Nonparametric Methods to Test Rare Variants for Multiple Traits

Xuexia Wang

University of Wisconsin-Milwaukee, United States. *xuexia@uwm.edu*

Abstract: Pleiotropy, effect of one variant on multiple traits, is a widespread phenomenon in complex diseases. Joint analysis of multiple traits such as systolic and diastolic blood pressures evaluated in hypertension can increase statistical power to detect disease susceptible genetic variants. Although the cost of next generation sequencing (NGS) has been reduced, it is still expensive to detect rare variants, and typically requires large samples. Moreover, most of the existing methods are parametric. These methods often assume parametric models and particular probability distributions for NGS data, and inference is made about the parameters of the distribution or model. However, for NGS data, those assumptions may not be correct and it is typically difficult or impossible to ascertain whether or not certain parametric assumptions are justifiable. Therefore, these parametric statistical methods can be very misleading. We develop new nonparametric statistical methods to efficiently detect rare variants for multiple traits, which can not only be applied in large samples, but also work for small samples. There are two steps in the nonparametric methods to detect rare variants. First, we rank the genotype data by mid ranks. Second, we weight the ranked genotype data with a quasi-optimal weight which will emphasize variants that have strong associations with the traits. Then, a nonparametric Dempster-ANOVA type statistic is applied. In order to evaluate the performance of the new methods, we conducted extensive simulation studies. For type I error evaluation, we considered different sample size, different haplotype structures, and different significance levels. In each simulation scenario, the estimated type I error rates of the new methods are under control. Thus, our new methods are valid tests to detect rare variants for multiple traits in next generation sequencing data. We will conduct more extensive simulation studies for power comparison with existing methods.

Thu, June 30 (13:30-15:10) l CP10

Probability Session 3

Chair: Zheng Zhang (Renmin University of China)

On Measure of Second-order Marginal Symmetry for Multi-way Contingency Tables

Yusuke Saigusa

Tokyo University of Science, Japan. *saigusaysk@gmail.com*

Abstract: Bhapkar and Darroch (1990) considered the second-order marginal symmetry model for multi-way contingency tables. If the goodness-of-fit of the second-order marginal symmetry model applied to the data is poorly, we may have interested in measuring the degree of departure from second-order marginal symmetry. So we shall propose the measure to represent degree of departure from second-order marginal symmetry. The proposed measure is expressed as the weighted sum of the Shannon entropy. The approximate confidence interval of the proposed measure is given. We shall show the proposed measure enable us to compare the degrees of departure from second-order marginal symmetry between two different tables.

Asymptotic Normality of Naive Canonical Correlation Coefficient in High Dimension Low Sample Size

Mitsuru Tamatani

Doshisha University, Japan. *mtamatan@mail.doshisha.ac.jp*

Abstract: In this talk we investigate the asymptotic behavior of the estimated naive canonical correlation coefficient under the normality assumption and high dimension low sample size settings. To avoid the singularity of usual sample covariance matrix in high dimension low sample size settings, we utilize the naive canonical correlation matrix with replacing sample covariance matrix by its diagonal part only. We derive the asymptotic normality of the estimated naive canonical correlation coefficient by using the central limit theorem for martingale difference sequence.

Generalized Asymmetry Models and Separations of Symmetry for Square Tables

Kouji Tahata

Tokyo University of Science, Japan. *kouji_tahata@is.noda.tus.ac.jp*

Abstract: We are interested in applying some kinds of models, which indicate the structure of symmetry (or asymmetry) rather than independence, for the analysis of square contingency tables with ordered categories. Bowker (1948) considered the symmetry model which indicates the structure that the cell probabilities are symmetric with respect to the main diagonal of the table. As the extension of the symmetry model, various types of symmetry (or asymmetry) models have been proposed. For example, quasi-symmetry (Caussinus, 1965), marginal homogeneity (Stuart, 1955), linear diagonals-parameter symmetry (Agresti, 1983). Also, Caussinus (1965) pointed out that the symmetry model is separated into the quasi-symmetry model and the marginal homogeneity model.

In this talk, we propose a model that indicates the structure of asymmetry for cell probabilities in the square contingency tables. The model indicates that the log-odds of symmetric cells are expressed as polynomial function of parameter, and includes some asymmetry models, which were proposed in early studies, in the special cases. Also, we show the theorem that the symmetry model can be separated into the proposed model and the model which indicates the structure of moment equality for marginal probabilities. Moreover, for the separation of symmetry, the relationships between test statistics are given. It may be useful to see the reason for the poor fit of the symmetry model on the details.

Log-normal Distribution Type Symmetry Model for Ordinal Square Contingency Tables

Kiyotaka Iki

Tokyo University of Science, Japan. *iki@is.noda.tus.ac.jp*

Abstract: For square contingency tables with the same row and column classifications, the symmetry model indicates that the symmetry of the probabilities with respect to the main diagonal of the table (Bowker, 1948). Agresti (1983) proposed the linear diagonals-parameter symmetry model which indicates that the log-odds that an observation will fall in the (i,j)th cell instead of in the (j,i)th cell, i<j, is proportional to the distance j-i from the main diagonal of the table. Agresti (1983) described the relationship between the linear diagonals-parameter symmetry model and the bivariate normal distribution with equal marginal variances. The linear diagonals-parameter symmetry model may be appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate normal distribution with equal marginal variances.

In this article, we propose new model which indicates that the log-ratios of symmetric cell probabilities are proportional to the difference between log-row category and log-column category. The proposed model may be appropriate for a square ordinal table if it is reasonable to assume an underlying bivariate log-normal distribution.

This article also gives a theorem such that the symmetry model holds if and only if both the proposed model and the log marginal mean equality model hold with the orthogonality of test statistics. When the symmetry model fits the data poorly, the theorem would be useful for seeing the reason for its poor fit. An example and simulation study are given.

Thu, June 30 (13:30-15:10) l CP16

Data Order Structure Session 3

Chair: Chun Yip Yau (The Chinese University of Hong Kong)

GARCH Modeling of Five Popular Commodities

Stephen Chan

The University of Manchester, United Kingdom. *stephen.chan@manchester.ac.uk*

Abstract: Flexible models for the innovation process of GARCH models have been limited.

Here, we show the flexibility of two recently proposed distributions due to Zhu and Zinde-Walsh (2009) and Zhu and Galbraith (2010) by means of GARCH modeling of five popular commodities. The five commodities considered are: Cocoa bean, Brent crude oil, West Texas intermediate crude oil, Gold and Silver. For each commodity, one of the two models due to Zhu and Zinde-Walsh (2009) and Zhu and Galbraith (2010) is shown to perform better than those commonly known.

Decomposing Time Series into Oscillation Components with Random Frequency Modulation

Takeru Matsuda

The University of Tokyo, Japan. *matsutake110@gmail.com*

Abstract: Many time series are naturally considered as a mixture of several oscillation components. For example, EEG time series are composed of several components such as alpha, beta and gamma. We propose a method for decomposing time series into such oscillation components with state space models. Based on the concept of random frequency modulation, Gaussian linear state space models for oscillation components are developed. In this model, the frequency of an oscillator is fluctuated by noise. Time series decomposition is accomplished by this model like the Bayesian seasonal adjustment method. Since parameters of the model are estimated from data by empirical Bayes method, the frequencies of oscillation components are determined in a data-driven manner. Also, appropriate number of oscillation components is determined with information criterion ABIC. In neuroscience, the phase of neural time series plays an important role in neural information processing. The proposed method can be used to estimate the phase of neural time series and it has several advantages over conventional method with the Hilbert transform.

A New Approach for Analyzing Panel AR(1) Series with Application to the Unit Root Test

Yu-Pin Hu

National Chi Nan University, Taiwan. *huyp@ncnu.edu.tw*

Abstract: This paper derives several novel tests to improve on the t-test for testing AR(1) co-efficients of panel time series, i.e., of multiple time series, when each having a small number of observations. These tests can determine the acceptance or the rejection of each hypothesis individually while controlling the average type one error. Strikingly, the testing statistics derived by the empirical Bayes approach can be approximated by a simple form similar to the t-statistic; the only difference is that the means and the variances are estimated by shrinkage estimators. Simulations demonstrate that the proposed tests have higher average power than the t-test in all settings we examine including those when the priors are miss-specified and the cross section series are dependent.

Detecting Differentially Methylated Regions via Non-homogeneous Hidden Markov Model

Linghao Shen

The Chinese University of Hong Kong, Hong Kong. *sl013@ie.cuhk.edu.hk*

Abstract: Methylation is a widely studied epigenetic marker, and the detection of differentially methylated regions is of great interest. Methylation data has special characteristics that are not utilized by existing methods. We propose a Non-homogeneous Hidden Markov Model to model the methylation data. The non-homogeneous transition model can meet the requirement to model the correlation with different probe distances. We also propose a special Gaussian Mixture Model to model the methylation values. Our approach allows us to model the methylation status and spatial correlation jointly. We compare our method with existing methods using both synthetic data and real data from TCGA, and in both cases our method outperformed exiting methods.

Bootstrap Method for Autoregressive Model

Bambang Suprihatin

University of Sriwijaya, Indonesia. *bambangs@unsri.ac.id*

Abstract: Usually we have data at hand with non standard assumptions, e.g. non-i.i.d. data, small data size, unknown distribution. In such situations, the bootstrap method is useful and can be applied. Bootstrap method also works well when applied to the time series data. However, the parameter estimates of the autoregressive model can be bootstrapped with accuracy that outperforms the normal approximation. We prove that the bootstrap parameter estimator of the autoregressive model converges in law to normal distribution.

Thu, June 30 (13:30-15:10) I CP17

Data Order Structure Session 4

Chair: Leming Qu (Boise State University)

Modified Kaplan-Meier Estimator and Nelson-Aalen Estimator with Geographical Weighting for Survival Data

Guanyu Hu

Florida State University, United States. *guanyu.hu@stat.fsu.edu*

Abstract: Kaplan-Meier estimator and Nelson-Aalen estimator, both non-parametric statistics, are universally used methods in the clinical studies. In the public health study, people often collect the data from different locations of the medical services provider. When some studies need to consider the survival curves from the different locations, the traditional estimators simply estimate the marginal survival curves by using stratification. In this paper, we used the idea from geographically weighted regression (Brunsdon et al (1996)) to add the geographical weight to the observations to get the modified Kaplan-Meier estimator and Nelson-Aalen estimator which can represent the local survival curves and cumulative hazard. We used the idea of counting process to get the modified estimator and the variance of the estimator. In addition, we introduced some general spatial weighting functions which can be applied into my modified estimators. Finally, we gave some simulation results of our estimators to illustrate the performance of the modified estimators.

Model Selection of Switching Mechanism for Financial Time Series

Chau Buu Truong

Feng Chia University, Taiwan. *buuchau@mail.fcu.edu.tw*

Abstract: The threshold autoregressive model with GARCH specification (TAR-GARCH) is a popular nonlinear model to capture the well-known asymmetric phenomena in financial market data. The switching mechanisms of hysteretic autoregressive GARCH (HAR-GARCH) models are different from TAR-GARCH in which the regime switching may be delayed when the hysteresis variable lies in a hysteresis zone. This paper investigates Bayesian model comparison among competing models by designing an adaptive Markov chain Monte Carlo (MCMC) sampling scheme. We illustrate the performance of three kinds of criteria in comparing models with fat-tailed and/or skewed errors: deviance information criteria (DIC), Bayesian predictive information, and its asymptotic version. A simulation experiment illustrates good performance in estimation and model selection. We demonstrate the proposed method in a study of 11 major stock markets.

Adjusted Kaplan-Meier Survival Curves for Marginal Treatment Effect in Observational Studies

Xiaofei Wang

Duke University School of Medicine, United States. *xiaofei.wang@duke.edu*

Abstract: For time-to-event outcome of multiple treatment groups, the Kaplan-Meier estimator is often used to estimate survival functions of treatment groups and compute marginal treatment effects. The Kaplan-Meier estimates and the derived estimates of marginal treatment effect are uniformly consistent under general conditions when data are from randomized clinical trials. For data from observational studies, however, these statistical quantities are often biased due to treatment-selection bias and the imbalanced distribution of baseline covariates that affect treatment assignment. Propensity score based methods, such as the inverse probability of treatment weighting, estimate the survival function by adjusting for the disparity of propensity scores between treatment groups. Unfortunately, the misspecification of the regression model will lead to biased estimates in these existing methods. Using an empirical likelihood (EL) method in which the moments of the covariate distribution of treatment groups are constrained to equality. After the empirical probability mass under the constraints are estimated, we obtain consistent estimates of the survival functions and the marginal treatment effect through the modified Kalpan-Meier estimator. Equating moments of the covariates distribution between treatments simulates the covariate distribution if the patients had been randomized to these treatment groups. We established the consistency and the asymptotic limiting distribution of the proposed EL estimators. We demonstrated that unlike propensity score methods, the consistency of the proposed estimator does not depend on a correct specification of a model. Simulation was used to study the finite sample properties of the proposed estimator and compare it with existing methods. The proposed estimator is illustrated with observational data from a lung cancer observational study to compare two surgical procedures in treating early stage lung cancer patients.

Quantile Regression Based on A Weighted Approach under Semi-competing Risks Data

Jin-Jian Hsieh

National Chung Cheng University, Taiwan. *jjhsieh@math.ccu.edu.tw*

Abstract: In this article, we investigate the quantile regression analysis for semi-competing risks data in which a non-terminal event may be dependently censored by a terminal event. Due to the dependent censoring, the estimation of quantile regression coefficients on the non-terminal event becomes difficult. In order to handle this problem, we assume Archimedean Copula (AC) to specify the dependence of the non-terminal event and the terminal event. Portnoy (2003) considered the quantile regression model under right censoring data. We extend his approach to construct a weight function, and then impose the weight function to estimate the quantile regression parameter for the non-terminal event under semi-competing risks data. We also prove the consistency and asymptotic properties for the proposed estimator. According to the simulation studies, the performance of our proposed method is good. We also apply our suggested approach to analyze a real data.

Regression Modeling of Interval Censored Competing Risk Data with Missing Cause

YangJin Kim

Sookmyung Women's University, South Korea. *yjin@sm.ac.kr*

Abstract: Several approaches have been suggested to analyze competing risk data in the presence of complete information of failure cause. However, missing causes for some subjects often occur. Furthermore, failure time also has interval censored form. For such incomplete information, we suggest a regression model using Andersen-Klein pseudo-value which is based on the estimated cumulative incidence function. Hudgens et al. suggested a nonparametric maximum likelihood for interval censored data with competing risk. By extending their methods, the regression coefficient is estimated using a multiple imputation and two methods are applied to estimate variance of the estimates. We evaluate the suggested method by comparing a complete case analysis in several simulation settings.

Thu, June 30 (15:30-17:10) | DL14 | Sponsor: Japan

Statistical Inference for Stochastic Processes: Asymptotic Theory and Implementation

Chair: Hiroki Masuda (Kyushu University)

Distinguished Lecturer: Nakahiro Yoshida (The University of Tokyo)

Statistics for Stochastic Processes: Inferential and Probabilistic Aspects

Nakahiro Yoshida

The University of Tokyo, Japan. *nakahiro@ms.u-tokyo.ac.jp*

Abstract: The quasi likelihood analysis and limit theorems give a mathematical foundation of statistical inference for stochastic processes.

The quasi likelihood analysis (QLA) is a framework of statistical inference, featuring the quasi likelihood random field and large deviation techniques. Limit theorems and tail probability estimates of the associated QLA estimators (QMLE, QBayesianE) follow systematically through QLA. This scheme applies to various dependence structures such as ergodic diffusion/jump-diffusion processes (Y AISM2011, Uchida-Y SPA2012, Ogihara-Y SISP2011). Estimation of volatility of a sampled semimartingale in finite time horizon is typical non-ergodic statistics. Asymptotic properties of the QLA estimators were proved in regular sampling (Uchida-Y SPA2013) and in non-synchronous sampling (Ogihara-Y SPA2014). Recently QLA was constructed for ergodic/non-ergodic point processes applied to limit order book (Ogihara-Y arXiv2015, Clinet-Y 2015).

Asymptotic expansion is a basis of various branches in statistics, e.g., higher-order efficiency, prediction, information criteria, resampling methods, information geometry, etc. After developments in mixing expansion and martingale expansion as refinements of CLT, asymptotic expansion in non-ergodic statistics has been an issue. Recently asymptotic expansion of a martingale with mixed normal limit was derived and applied to the realized volatility (Y SPA2013). The Malliavin calculus (infinite-dimensional stochastic calculus) plays an essential role there. The second-order approximation was applied to derive a spot volatility information criterion (sVIC) for volatility model selection (Uchida-Y 2015). The martingale expansion method also applies to the QLA estimators of volatility parameters.

The formulas obtained by inferential theory are often quite complicated. YUIMA, an R framework for statistical inference and simulation for stochastic processes, is incorporating the latest theoretical results.

Hybrid Type Estimation for Diffusion Type Processes Based on High Frequency Data

Masayuki Uchida

Osaka University, Japan. *uchida@sigmath.es.osaka-u.ac.jp*

Abstract: We consider the estimation problem of the unknown parameter for diffusion type processes based on discrete observations. For both ergodic and non-ergodic diffusion processes, some asymptotic properties including the convergence of moments for the maximum likelihood (ML) type estimator and the Bayes type estimator have been shown, see Uchida and Yoshida (2012, SPA; 2013, SPA; 2014, SISP). From the viewpoint of numerical analysis, however, the optimization of the quasi likelihood function causes a serious problem since the quasi likelihood functions of diffusion type processes generally have complex forms. Furthermore it takes a lot of time to compute the Bayes type estimators. Recently, for ergodic diffusion processes, Kamatani and Uchida (2015, SISP) proposed the hybrid multi-step estimator with the asymptotic normality and the convergence of moments. Here we note that our method works well even if an initial estimator does not have the optimal rate of convergence. In this talk, we apply our method to the volatility estimation for non-ergodic diffusion type processes. First, we obtain the initial Bayes type estimator with non-optimal rate of convergence, and then calculate the hybrid multi-step estimator by means of Newton-Raphson method with the initial Bayes type estimator. We show that the multi-step estimator has asymptotic mixed normality with convergence of moments. Moreover, an example and simulation results are given. This is a joint work with Kengo Kamatani and Akihiro Nogita.

Computational Aspects of Simulation and Inference for CARMA and COGARCH Models

Stefano Iacus

The University of Milan, Italy. *stefano.iacus@unimi.it*

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 Lévy model for both estimation and simulation. The estimation of the parameters for the underlying Lévy 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 Lévy process is used to increase the accuracy of the estimates through aggregation of the increments.

Thu, June 30 (15:30-17:10) I IP17 I Sponsor: Bernoulli Society

Recent Advances and Trends in Time Series Analysis

Organizer: Liudas Giraitis (Queen Mary University of London)

Chair: Liudas Giraitis (Queen Mary University of London)

On Consistency/Inconsistency of MDL Model Selection for Piecewise Autoregressions

Richard A. Davis

Columbia University, United States. *davis.richarda@gmail.com*

Abstract: The Auto-PARM (Automatic Piecewise AutoRegressive Modeling) procedure, developed by Davis, Lee, and Rodriguez-Yam (2006), uses the minimum description length (MDL) principle to estimate the number and locations of structural breaks in a non-stationary time series. Consistency of this model selection procedure has been established when using conditional maximum (Gaussian) likelihood variance estimates. In contrast, the estimate of the number of change-points is inconsistent in general if Yule-Walker variance estimates are used instead. This surprising result is due to an exact cancellation of first-order terms in a Taylor series expansion in the conditional maximum likelihood case, which does not occur in the Yule-Walker case. (This is joint work with Stacey Hancock and Yi-Ching Yao.)

Structure Identification in Panel Data Analysis

Wenyang Zhang

University of York, United Kingdom. *wenyang.zhang@york.ac.uk*

Abstract: Panel data analysis is an important topic in statistics and econometrics. It is very common to assume the impact of a covariate on the response variable remains constant across all individuals. In general, this assumption may overlook some individual/subgroup attributes of the true covariate impact.

In this talk, I will show a data driven approach to identify the groups in panel data with interactive effects induced by latent variables. An EM based algorithm is proposed to estimate the unknown parameters, and a binary segmentation based algorithm is proposed to detect the grouping. I will also show some asymptotic theories and simulation studies to justify the proposed methods. Finally, I will show a real data analysis to illustrate the proposed methods.

Root-n Consistent Estimation of the Marginal Density of Some Stationary Time Series

Lionel Truquet

ENSAI, France. *lionel.truquet@ensai.fr*

Abstract: It is well known that, under some conditions, the density of a function of several independent random variables can be estimated at the usual parametric rate of convergence, using U-statistics arguments and kernel density estimation. For regression models and time series models, some results have recently been obtained.

In this talk, we will first make a review of the available results and next, we will give some extensions to conditionally heteroscedastic time series models such as GARCH processes.

Inference for Conditionally Heteroscedastic Location-scale Time Series Models

Sangyeol Lee

Seoul National University, South Korea. *jpgrslee@gmail.com*

Abstract: In this study, we investigate the asymptotic properties of conditionally heteroscedastic location-scale time series models with innovations following a generalized asymmetric Student-t distribution (ASTD) and asymmetric exponential power distribution (AEPD). We show that the MLE is consistent and asymptotically normal under some conditions, and use the MLE to estimate the conditional Value-at-Risk (VaR) and Expected Shortfall (ES). A comparison study with CAViaR and CARE methods is conducted. Further, to check the model adequacy, we consider the entropy based goodness of fit test and the residual based CUSUM test. As an illustration, a real data analysis is provided.

Thu, June 30 (15:30-17:10) l IP39 l Sponsor: Australia and New Zealand

Advances in Statistical Inference for Multivariate Response Data

Organizer: Samuel Mueller (University of Sydney)

Chair: Samuel Mueller (University of Sydney)

Vector Regression Without Marginal Distributions or Association Structures

Alan Huang

The University of Queensland, Australia. *alan.huang@uq.edu.au*

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.

Estimation and Inference in Directional Mixed Models for Compositional Data

Janice Lea Scealy

Australian National University, Australia. *janice.scealy@anu.edu.au*

Abstract: Compositional data are vectors of proportions defined on the unit simplex and this type of constrained data occur frequently in applications. It is also possible for the compositional data to be correlated due to the clustering or grouping of the observations. We propose a new class of mixed model for compositional data based on the Kent distribution for directional data, where the random effects also have Kent distributions. The advantage of this approach is that it handles zero components directly and the new model has a fully flexible underlying covariance structure. One useful property of the new directional mixed model is that the marginal mean direction has a closed form and is interpretable. The random effects enter the model in a multiplicative way via the product of a set of rotation matrices and the conditional mean direction is a random rotation of the marginal mean direction. For estimation we apply a quasi-likelihood method which results in solving a new set of generalised estimating equations and these are shown to have low bias in typical situations. For inference we use a nonparametric bootstrap method for clustered data which does not rely on estimates of the shape parameters (shape parameters are difficult to estimate in Kent models). The new approach is shown to be more tractable than the traditional approach based on the logratio transformation.

Bayesian Gegenbauer Long Memory Financial Time Series Models

Jennifer Chan

The University of Sydney, Australia. *jennifer.chan@sydney.edu.au*

Abstract: We discuss a time series model that is generalized long memory in the mean process. We develop a new Bayesian posterior simulator that couples advanced posterior maximisation techniques. Details are provided on the estimation process, data simulation, and out of sample performance measures. We conduct several rigorous simulation studies which verify our results in and out of sample. Further, we provide an empirical in-sample application with out-of-sample forecast and compare performance of the generalized process to the usual long memory filter.

Stability Selection with Information Criteria

Zhen Pang

The Hong Kong Polytechnic University, Hong Kong. *zhen.pang@polyu.edu.hk*

Abstract: A popular approach to model high dimensional data is to use the regularization penalty. However, the final selected model is determined by a tuning parameter which controls the level of penalty. Thus, it is a fundamental problem to determine this tuning parameter. We propose a novel tuning parameter selection method by combining information criteria and stability selection to enjoy the advantages of the two methods while avoiding their drawbacks. The new method is quite general and can be generalized to many classes of models like the generalized linear models, Cox's proportional hazard model in survival analysis, graphical models and clustering.

Thu, June 30 (15:30-17:10) | IP50

Semiparametric Statistical Methods for Complex Problems

Organizer: Lan Wang (University of Minnesota)

Chair: Lan Wang (University of Minnesota)

Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection

Runze Li

The Pennsylvania State University, United States. *rzli@psu.edu*

Abstract: Due to the need of studying the theoretical property of variable selection procedure for Cox's model, we study the asymptotic behavior of partial likelihood for the Cox model. We find that the partial likelihood does not behave like an ordinary likelihood, whose sample average typically tends to its expected value, a finite number, in probability. Under some mild conditions, we prove that the sample average of partial likelihood tends to infinity at the rate of logarithm of the sample size in probability. This is an interesting and surprising results because the maximum partial likelihood estimate has the same asymptotical behavior as the ordinal maximum likelihood estimate. We further apply the asymptotic results on the partial likelihood to study tuning parameter selection for penalized partial likelihood. Our finding indicates that the penalized partial likelihood with the generalized cross-validation (GCV) tuning parameter proposed in Fan and Li (2002) enjoys the model selection consistency property. This is another surprising result because it is well known that the GCV, AIC and Cp are all equivalent in the context for linear regression models, and are not model selection consistent. Our empirical studies via Monte Carlo simulation and real data example confirm our theoretical finding.

Modeling Complex and Big Survival Data: Computation and More

Yi Li

University of Michigan, United States. *yili@umich.edu*

Abstract: Modern data collection techniques have resulted in an increasing number of big clustered time-to-event data sets, wherein patients are often observed from a large number of clusters, e.g., healthcare providers. Semiparametric frailty models are a flexible and powerful tool for modeling clustered time-to-event data. However, the computational complexity of frailty models has limited their use in big time-to-event data. When the sample size or the number of parameters is large, many existing methods fail because of lack of computational power. In this manuscript, we provide a computationally efficient approach based on a minimization-maximization algorithm to fit semiparametric frailty models in large-scale settings. Moreover, the proposed method is extended to incorporate complex data structures such as time varying effects. The utility and finite-sample properties are examined through an extensive simulation study, and an analysis of national kidney transplant data.

Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference

Ingrid van Keilegom

Université catholique de Louvain, Belgium. *ingrid.vankeilegom@uclouvain.be*

Abstract: In both parametric and certain nonparametric statistical models, the empirical likelihood ratio satisfies a nonparametric version of Wilks' theorem. For many semiparametric models, however, the commonly used two-step (plug-in) empirical likelihood ratio is not asymptotically distribution-free, that is, Wilks' phenomenon breaks down. In this paper we suggest a general approach to restore Wilks' phenomenon in two-step semiparametric empirical likelihood inferences. The main insight consists in using as the moment function in the estimating equation the influence function of the plug-in sample moment. The proposed method is general, leads to distribution-free inference and it is less sensitive to the first-step estimator than alternative bootstrap methods. Several examples and a simulation study illustrate the generality of the procedure and its good finite sample performance. (This is joint work with Juan Carlos Escanciano and Francesco Bravo.)

Semi-nonparametric Inference in Possibly Misspecified Regression Models with Missing Data

Phillip Sheung Chi Yam

The Chinese University of Hong Kong, Hong Kong. *scpyam@sta.cuhk.edu.hk*

Abstract: Regression analysis with missing outcome or covariates has been studied under various parametric and semiparametric framework, where modelling of missing data mechanism or covariate distribution other than the regression model of primary interest is usually inevitable. Those additional models, together with the regression model of interest, can all be misspecified in practice and it is difficult to even understand what parameters are being estimated if the models are misspecified. We consider parameters that are defined transparently if the regression model is misspecified, and study a general class of nonparametric calibration weighting method that can attain efficient estimation under a missing-at-random assumption, without resorting to additional models or nonparametric estimation of known distributions which are otherwise needed in existing methods. A consistent estimator for the efficient asymptotic variance is also proposed without additional estimation efforts, and inference based on a large-sample approximation is accurate in practical sample sizes.

Thu, June 30 (15:30-17:10) l IP58

Stochastic Partial Differential Equations

Organizer: Jian Song (The University of Hong Kong)

Chair: Jian Song (The University of Hong Kong)

Density of Parabolic Anderson Field

Yaozhong Hu

The University of Kansas, United States. *yhu@ku.edu*

Abstract: Let $u(t,x)$ be the solution of the parabolic Anderson equation. Denote by $\varrho(t,x;y)$ the density with respect to the Lebesgue measure of the law of $u(t,x)$. The smoothness of $\varrho(t,x;y)$ as a function of y is known. We want to find the lower and upper bound of this density function $\varrho(t,x;y)$ as $y]\to\infty$ and as $y\to 0$.

Poincare Inequality for Dirichlet Distributions and Infinite-dimensional Generalizations

Feng-Yu Wang

Beijing Normal University, China. *wangfy@bnu.edu.cn*

Abstract: The Dirichlet distribution and its infinite-dimensional generalization arise naturally in Bayesian inference as conjugate priors for categorical distribution and infinite non-parametric discrete distributions respectively. They also arise in population genetics describing the distribution of allelic frequencies. By establishing the sharp Poincare inequality, the exact exponential convergence rate is found for a class of diffusion processes to converge to the Dirichlet distribution. Moreover, the sharp Poincare inequality is extended to the infinite-dimensional setting, and the spectral gap of the corresponding discrete model is derived.

Large Deviation Principle for the Occupation Measures of Stochastic PDEs

Lihu Xu

University of Macau, Macao. *lihuxu@umac.mo*

Abstract: Using the hyper-exponential recurrence criterion, a large deviation principle for the occupation measure is derived for a class of non-linear monotone stochastic partial differential equations (SPDEs). The main results are applied to many concrete SPDEs such as stochastic p-Laplace equation, stochastic porous medium equation, stochastic fast-diffusion equation, and even stochastic Ginzburg-Landau equation driven by heavy tailed Levy noises. This is the joint work with Ran Wang and Jie Xiong.

Thu, June 30 (15:30-17:10) l CP07

Bayesian Session 2

Chair: Tianwei Yu (Emory University)

New Method for Revealing Free Energy Landscape of Proteins

Hangjin Jiang

The Chinese University of Hong Kong, Hong Kong. *cas.jiang@gmail.com*

Abstract: Molecular dynamics simulation provides rich data for studying free energy landscape, we proposed a new method called Conditional Angle Partition Tree(CAPT) to reveal protein free energy landscape. CAPT is an efficient method comparing with existed methods, such as PCCA, PCCA+, and MPP(Most Probable Pathway), which is illustrated through a benchmark molecular dynamics data of Ala-dipeptide. Different from PCCA, PCCA+ and MPP, CAPT does not need to cluster firstly frames into microstates. Instead, it performs on frames, and considers relationship between geometrical similarity and dynamical similarity, which is ignored by other methods. In addition, the global geometrical similarity is gradually fulfilled. This is a joint work with Xiaodan Fan.

A Fast and Powerful W-test for Pairwise Gene-gene Interaction Testing in GWAS Data

Maggie Wang

The Chinese University of Hong Kong, Hong Kong. *maggiew@cuhk.edu.hk*

Abstract: Epistasis plays an essential role in the development of complex diseases. Interaction methods face the common challenge of seeking a balance between persistent power, model complexity, computation efficiency, and validity of identified bio-markers. We introduce a novel W-test to identify pairwise epistasis effect, which measures the distributional difference between cases and controls through a combined log odds ratio. The test is model-free, fast, and inherits a Chi-squared distribution with data adaptive degrees of freedom. No permutation is needed to obtain the p-values. Simulation studies demonstrated that the W-test is more powerful than alternative methods in low frequency variants environments, which are the Chi-squared test, logistic regression and multifactor-dimensionality reduction (MDR). In two independent real bipolar disorder genome-wide associations (GWAS) datasets, the W-test identified significant interactions pairs that can be replicated, including SLIT3-CENPN, SLIT3-TMEM132D, CNTNAP2-NDST4 and CNTCAP2-RTN4R. The genes in the pairs play central roles in neurotransmission and synapse formation. A majority of the identified loci are undiscoverable by main effect and are low frequency variants. The proposed method offers a powerful alternative tool for mapping the genetic puzzle underlying complex disorders.

Closed Form Bayesian Inferences for Binary Logistic Regression

Kevin Dayaratna

The Heritage Foundation, United States. *kevin.dayaratna@heritage.org*

Abstract: In many fields, logistic regression has become one of the most widely used tools in statistical modeling. With improvements in computing power over the past two decades, incorporating heterogeneity in these and other models has become increasingly common in the statistical literature. Parametric Bayesian models are often used to incorporate individual-level heterogeneity. Although "nice" in principle, incorporating individual-level heterogeneity is often concomitant with the drawbacks of the computational complexity associated with numerical computation, especially for large data sets involving high-dimensional parameter spaces.

In this research, we present an alternative estimation technique for Bayesian binary logistic regression using polynomial expansions. We present a series of simulations as well as an application to child poverty to illustrate the efficacy of our approach. We find that this approach significantly outperforms existing Bayesian estimation methods in terms of computing time. These gains are extremely useful for large data sets.

Thu, June 30 (15:30-17:10) I CP05

Machine Learning Session 5

Chair: Hironori Fujisawa (The Institute of Statistical Mathematics)

A One-stage Approach for Principal Component Regression via L1-type Regularization

Shuichi Kawano

The University of Electro-Communications, Japan. *skawano@ai.is.uec.ac.jp*

Abstract: Principal component regression (PCR) is widely used in various fields of research; chemometrics, bioinformatics, and so on. As PCR selects some principal components and then constructs a regression model regarding them as new explanatory variables, it is a two-stage approach. This two-stage approach, however, causes some problems. For example, as the principal components are obtained from only explanatory variables and not considered with the response variable, the PCR may not have enough prediction accuracy. To overcome the problems, we propose a one-stage approach for PCR along with the technique of L1-type regularization. We call this method a sparse principal component regression (SPCR). SPCR enables us to obtain principal component loadings that are related to the response variable and select the number of principal components. To derive the estimates of parameters in SPCR, we employ the coordinated descent algorithm, since SPCR is the convex optimization problem for each of parameters with sparse regularization terms. Some numerical examples are given to illustrate the effectiveness of SPCR. The R language software package spcr, which implements SPCR, is available from the Comprehensive R Archive Network (CRAN).

Generalized Principal Component Analysis: Dimensionality Reduction through the Projection of Natural Parameters

Yoonkyung Lee

The Ohio State University, United States. *yklee@stat.osu.edu*

Abstract: Principal component analysis (PCA) is useful for a wide range of data analysis tasks. However, its implicit link to the Gaussian distribution can be undesirable for discrete data such as binary and multi-category responses or counts. We generalize PCA to handle various types of data using the generalized linear model framework. In contrast to the existing approach of matrix factorizations for exponential family data, our generalized PCA provides low-rank estimates of the natural parameters by projecting the saturated model parameters. Due to this difference, the number of parameters does not grow with the number of observations and the principal component scores on new data can be computed with simple matrix multiplication. We provide a computationally efficient algorithm for finding the principal component loadings and demonstrate the benefits of the proposed approach numerically.

This is joint work with Andrew Landgraf.

A High Dimensional Two-sample Test using Nearest Neighbors Based on a New Dissimilarity Measure

Rahul Biswas

Indian Statistical Institute, India. *1992.rahul@gmail.com*

Abstract: Schilling (1986) and Henze (1988) proposed a nonparametric two-sample test using nearest neighbors based on Euclidean distances. Though this test is consistent in classical asymptotic regime, it often performs poorly for high dimensional data. In high dimensions, phenomena like concentration of pairwise distances and violation of class assumptions often have adverse effects on its performance, especially when the scale difference between two distributions dominates the difference in their locations. In this project, we propose a nonparametric two-sample test based on nearest neighbor type coincidences, where instead of Euclidean distance, a new dissimilarity measure is used. This dissimilarity measure is based on average of absolute differences between inter-point distances, and it uses the distance concentration property to its advantage to yield better performance in high dimensions. While several existing tests for the multivariate two- sample problem are not applicable when the dimension exceeds the sample size, our proposed test can be conveniently used in the high dimension low sample size (HDLSS) situations. Unlike Schilling and Henze's test, under fairly general conditions, this new test is found to be consistent in HDLSS asymptotic regime, where the sample size remains fixed and the dimension grows to infinity. We analyze several high dimensional simulated and real data sets to compare their empirical performance with some popular two-sample tests available in the literature.

Thu, June 30 (15:30-17:10) l CP13

Big Data/Network Session 3

Chair: Zheng Zhang (Renmin University of China)

A Dynamic Logistic Regression for Network Link Prediction

Jing Zhou

Peking University, China. *jing.zhou@pku.edu.cn*

Abstract: In social network analysis, link prediction has gained increasing attention. How to conduct a comprehensive and principled evaluation, about various structure information in dynamic social network, is a problem of interest. To this end, we propose here a dynamic logistic regression method. Specifically, we assume that one has observed a time series of network structure. Then the proposed model dynamically predicts future links by studying historical changes of network structure. In particular, we propose three important effects to capture the network characteristics: momentum effect, homophily effect and common factor effect. The network structure is allowed to be extremely sparse. To estimate the model, we find that the standard maximum likelihood (MLE) is computationally forbidden. To solve the problem, we introduce a novel conditional likelihood (CMLE) method, which is computationally feasible for large-scale networks. This approach balances the limited amount of available information and the huge amount of prediction complexity in large-scale dynamic social networks. We demonstrate the performance of the proposed method with simulation studies and a real data example.

Promote Gross Capital Formation with Internet Access

Aulia Dini

STIS 53 Economics Division, Indonesia. *auliadini.stis53@gmail.com*

Abstract: Internet Access Survey, has an establishment approach, is conducted in 29 districts and 46 big cities(including 33 regional capitals) with exact building and street address. Household Investment Survey of 4500 households is based on statistical area without building address and without street address. Strongest signal is assumed if a household is located within a vicinity of signal available both to business and educational institution. Strong signal is assumed if the vicinity has only signal available to business. With almost fifty exact matched households in districts median gross capital formation is six million Rupiah and highest gross capital formation is 300 million Rupiah. 150 households in regional capital and regional second biggest city have median gross capital formation of five million Rupiah and highest gross capital formation of almost 600 million Rupiah. A chance exists in districts to promote gross capital formation approaching that of regional second biggest city and regional capital hopefully with internet help. Since gross capital formation is derived then there is a chance to promote other variates including fixed capital, monetary transfer, stock dividend.

A Cognitive Model of Data Visualization in Big Data Analytics

Ken W Li

Hong Kong Institute of Vocational Education, Hong Kong. *kenli@vtc.edu.hk*

Abstract: Data can be easily gathered via the Internet and processed and managed efficiently by technology. The sheer volume and complexity data can be analyzed by big data analytics tools. Data visualization is one the tools which provides a fast way of extracting useful information from big data but many students or business analysts have difficulty in gaining insights and seeing hidden relationships of data. This hinders their ability of derive intrinsic meaning of data. In this paper, the author presents a critical review of current research in this area from three different perspectives: pedagogy, statistics and cognitive psychology. Arising from a synthesis of this research he proposes a cognitive model of data visualization which helps them to check whether data form patterns; search for data patterns; check whether data patterns have meaning; and interpret the meaning of data patterns.

The Estimation of Claims Reserve with Reserving by Detailed Conditioning Method (RDC)

Adhitya Ronnie Effendie

Gadjah Mada University, Indonesia. *adhityaronnie@yahoo.com*

Abstract: We will estimate claims reserve by using Reserving by Detailed Conditioning (RDC) method. RDC method is an estimation method of claim reserve which involves claims information (claims characteristics) in the calculation process. It uses claims data which are represented binto individual run-off triangle as input data. The estimation of claim reserve will be applied to the liability insurance's claim data. The result of this estimation will be compared with the result of Chain Ladder (CL) method. As the result of this research, the calculation of claim reserve estimation using RDC method gives fewer MSEP value than CL method does.

Thu, June 30 (15:30-17:10) l CP18

Data Order Structure Session 5

Chair: Yu-Pin Hu (National Chi Nan University)

Comparing Predictive Values of Two Diagnostic Tests in Small Clinical Trials

Kouji Yamamoto

Osaka University, Japan. *yamamoto-k@stat.med.osaka-u.ac.jp*

Abstract: In screening tests or diagnostic tests, there are four important measures for quantifying test accuracy. Sensitivity is the probability of test outcome being positive among diseased subjects, and specificity is the probability of test outcome being negative among non-diseased subjects. On the other hand, positive predictive value (PPV) is the probability of a subject being diseased when the test outcome is positive, and the negative predictive value (NPV) is the probaility of a subject not being diseased when the test outcome is negative. When we want to compare the sensitivities (or specificities) for two tests evaluated same subjects, McNemar's test is frequently used. However, there are few statistical researches (methodologies) for comparing the PPVs (or NPVs), though several authors discussed that the PPVs (or NPVs) were so important for patient care. In addition, for the comparison of the predictive values, all of existing methods were argued in moderate or large sample situation. In this presentation, we mention the performance of existing methods for comparing the predictive values when the sample size is small, and describe an improvement of estimates for the difference of PPVs (or NPVs).

On Coding and Centering in the Autologistic Regression Model

Mark Wolters

Fudan University, China. *mwolters@fudan.edu.cn*

Abstract: Autologistic regression is useful for modelling binary responses with a complex association structure, as can arise in spatial analyses or image processing. It is standard in the statistical literature to code binary variables as 0 and 1, but with this coding the model parameters are hard to interpret. A centered version of the model has been proposed by Caragea and Kaiser (2009) to address this problem. At the same time, in other fields it is more common to see a -1, +1 coding used with the autologistic model, without the centering adjustment. This talk first demonstrates that the different combinations of coding and centering are not trivial transformations of the same model: they in fact represent different probabilistic structures. Some demonstrations of the differences will be given, and it will be argued that the plus/minus coding with no centering adjustment--not the centered 0/1 model--is the best choice for most situations.

Automatic Optimal Batch Size Selection for Recursive Estimators of Time-average Covariance Matrix

Kin Wai Chan

Harvard University, United States. *kinwaichan@g.harvard.edu*

Abstract: The time average covariance matrix (TACM) $\Sigma = \sum_{k \in Z} \Gamma_k$, where Γ_k is the auto-covariance function, is an important quantity for the inference of the mean of a R^d-valued stationary process ($d \geq 1$). This paper proposes two recursive estimators for Σ with optimal asymptotic mean square error (AMSE) under different strengths of serial dependence. The optimal estimator involves a batch size selection, which requires knowledge of a smoothness parameter $\Upsilon_\beta = \sum_{k \in Z} |k|^\beta \Gamma_k$, for some β. This paper also develops recursive estimators for Υ_β. Combining these two estimators, we obtain a fully automatic procedure for optimal on-line estimation for TACM. Consistency and convergence rates of the proposed estimators are derived. Applications to confidence region construction and Markov Chain Monte Carlo convergence diagnosis are discussed.

Covariate-adjusted Response-adaptive Designs for Weibull Survival Responses

Ayon Mukherjee

Queen Mary University of London, United Kingdom. *atg.dcb@gmail.com*

Abstract: Covariate-adjusted response-adaptive (CARA) designs use the available responses to skew treatment allocation in a clinical trial in favor of the treatment found at an interim stage of the trial to be best for a given patient's covariate profile. There has recently been extensive research on diverse aspects of CARA design which mainly involved binary response trials. Though exponential survival responses have also been considered, their constant hazard property makes the mean residual life for patients constant. This model is thus too restrictive for wide-ranging applications. To address this shortcoming, the proposed designs have been developed for Weibull survival responses by deriving two variants of the optimum designs and by using a link-function. The optimal designs are based on the doubly adaptive biased coin design (DBCD), and the efficient randomized adaptive design (ERADE). The allocation proportion to a treatment for these designs converges to the expected targeted values. The use of a link function would be appropriate in the context of a Weibull model when survival times are right censored. Such a link function based on the cumulative distribution function of a Gumbel distribution has been derived. An expression for the conditional probability that a given patient will be allocated to a particular treatment has been obtained. The ERADE is preferable to the DBCD if minimizing the variance of the allocation procedure is the main aim. However, the former procedure being discrete will tend to be slower in converging towards the expected target allocation proportion. The optimal allocation approach is better than the link function one as far as the power of Wald test is concerned. If the ethical objective is the sole consideration the link function based procedure would be preferable. An extensive simulation study of the operating characteristics of the proposed designs supports these findings.

Thu, June 30 (13:30-15:10) | CP19

Data Order Structure Session 6

Chair: YangJin Kim (Sookmyung Women's University)

The Empirical Beta Copula

Hideatsu Tsukahara

Seijo University, Japan. *tsukahar@seijo.ac.jp*

Abstract: Given a sample from a multivariate distribution *F*, the uniform random variates generated independently and rearranged in the order specified by the vector of ranks look like a sample from the copula of *F*. This idea can be regarded as Baker's construction of copulas based on order statistics with the ranks being coefficients, and led us to define the empirical beta copula. It is then fairly easy to show that the empirical beta copula is a particular case of the empirical Bernstein copula (taking all the orders of the Bernstein polynomials equal to the sample size). The advantage is that we do not need any smoothing parameter. Also it is extremely simple to simulate a sample from the empirical beta copula. We show that the empirical Bernstein copula is a genuine copula by providing a necessary and sufficient condition for a Bernstein transformation to be a copula. Furthermore, we establish the assumptions under which the standard asymptotic results hold for the empirical Bernstein copula. They are significantly weaker than those given in Janssen et al. Our Monte Carlo simulation study shows that there is an advantage of smoothing to improve finite-samples performance. It is found that in all cases, the empirical beta copula outperforms the empirical copula in terms of the bias and the integrated mean squared error. Compared with the empirical Bernstein copula with optimal smoothing rate, its performance is still significantly better in several cases, especially in terms of bias.

Multivariate Copula Density Estimation by Mixture of Parametric Copula Densities

Leming Qu

Boise State University, United States. *lqu@boisestate.edu*

Abstract: As a stochastic dependence modeling tool beyond the classical normal distribution model, Copula is widely used in Economics, Finance, and Engineering. A Multivariate copula density estimation method that is based on finite mixture of parametric Copula densities is proposed here. More specifically, one component of the mixture model is the mixture of Gaussian, Clayton and Gumbel Copulas (termed GCG component) which are capable of capturing symmetrical, lower, and upper tail dependence, respectively. The entire copula density is a mixture of k GCG components. The dimensionality of the density could be higher than 2. For dimensions higher than 2, restricted correlation matrix for the Gaussian components reduces the number of unknown parameters. The model parameters are estimated by interior-point algorithm for the resulting constrained maximum likelihood estimation problem, where the gradient of the objective function is not required. The interior-point algorithm is compared with the commonly used expectation maximization (EM) algorithm in mixture models. Mixture components with small weights can be removed by a thresholding rule. The number of components k is selected by the model selection criterion AIC. Simulation and real data application show the effectiveness of the proposed approach.

Nonparametric Wind Power Forecasting under Fixed and Random Censoring

Georgios Effraimidis

University of Southern California, United States. *gef@sam.sdu.dk*

Abstract: We consider nonparametric forecasting of wind power for individual wind turbines, allowing for random right censoring as well as two-sided fixed censoring. We propose a very fast estimation algorithm and show that this estimator of the unknown regression function is uniformly consistent. We argue that the key statistical features of the proposed nonparametric regression framework such as nonlinearities, fixed and random censoring are all needed in order to properly capture the main characteristics of wind power production functions. We provide an empirical illustration comparing forecast accuracy of the proposed nonparametric regression model to some of the existing and popular forecasting devices in predicting short to medium term wind power production at the individual turbine level. The empirical results are generally very encouraging.

Dynamic Prediction of Alzheimer's Disease Risk Based on Longitudinal Biomarkers and Functional Data

Sheng Luo

University of Texas at Houston, United States. *sheng.t.luo@uth.tmc.edu*

Abstract: In the study of Alzheimer's disease (AD), an important survival outcome is the progression from mild cognitive impairment (MCI) to AD. An accurate prediction of the time from MCI to AD is helpful for physicians to monitor patients' disease progression and make informative medical decisions. We propose a dynamic prediction framework based on a joint model that consists of a longitudinal regression model with functional exposure (high dimensional magnetic resonance imaging) and a survival model for event time. This framework provides accurate prediction of target patients' future outcome trajectories and risk of AD conversion. Our proposed model is motivated and applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI-1).

Poster Session

Poster Presentation

Birth Cohort Effect in Japan - Automatic Detection and Statistical Evaluation (PS02)

Tetsuji Tonda

Prefectural University of Hiroshima, Japan. *ttetsuji@pu-hiroshima.ac.jp*

Abstract:
Background: Cancer mortality is increasing with the aging of the population in Japan. Cancer information obtained through feasible methods is therefore becoming the basis for planning effective cancer control programs. There are three time-related factors affecting cancer mortality, of which the cohort effect is one. Past descriptive epidemiologic studies suggest that the cohort effect is not negligible in cancer mortality.

Methods: In this paper, we develop a statistical method for automatically detecting a cohort effect and assessing its statistical significance for cancer mortality data using a varying coefficient model.

Results: The proposed method was applied to liver and lung cancer mortality data on Japanese men for illustration. Our method detected significant positive or negative cohort effects. The relative risk was 1.54 for liver cancer mortality in the cohort born around 1934 and 0.83 for lung cancer in the cohort born around 1939.

Conclusions: Cohort effects detected using the proposed method agree well with previous descriptive epidemiologic findings. In addition, the proposed method is expected to be sensitive enough to detect smaller, previously undetected birth cohort effects.

High Dimensional LASSO Variable Selection Under Strongly-correlated Covariates (PS03)

Kaimeng Zhang

Chonnam National University, South Korea. *18639260887@163.com*

Abstract: In this talk, we consider a regression model that regress the response against the idiosyncratic factors obtained from the factor analysis of the covariates. Such a model is particularly useful in the high-dimensional variable selection problems in certain econometrics applications where all covariates are correlated due to systematic economic factors. In such cases, it is shown both theoretically and empirically that the usual penalized regression of the response against the covariates tends to select either all or none of the covariates. On the contrary, the proposed hybrid approach of factor analysis and penalized regression can select relevant covariates consistently under p=O(e^n) and other mild conditions on the factor loading matrix, where p and n are the number of covariates and the sample size respectively. To illustrate the ideas, two empirical data analysis examples are considered, (i) the gross domestic production of a chosen country against capital inputs and labor inputs of all countries and (ii) the stock returns of a chosen stock against the trading volumes of all stocks in the financial market.

Growth Curve Model with Nonparametric Baselines and Its Statistical Inference (PS04)

Kenichi Satoh

Hiroshima University, Japan. *ksatoh@hiroshima-u.ac.jp*

Abstract: This paper presents a method for estimating the regression coefficients for a growth curve model when the time trend of the baseline has not been specified. The concept of this method is similar to that of the Cox proportional hazard model. No particular shape is assumed for the baseline time trends, or, alternatively, it can be assumed that they are estimated nonparametrically. Due to these nuisance parameters for the baseline trends, we do not have to pay attention to model those shapes. Besides the simplicity of modeling baseline curves, we can also nonparametrically describe the baseline trends by using the residuals after the regression coefficients have been estimated.

Evaluating Standard Errors of Total Heritability Estimate in Genome-wide Association Studies Based on Summary Statistics Alone (PS05)

Hon-Cheong So

The Chinese University of Hong Kong, Hong Kong. *hcso85@gmail.com*

Abstract: Genome-wide association studies (GWAS) have become increasing popular these days and one of the key questions is how much heritability can be explained by all variants in GWAS. We have previously devised an approach to answer this question, based on recovering the "true" z-statistics from a set of observed z-statistics. Only summary statistics are required. However, methods for standard error (SE) estimation are not available yet, limiting the interpretation of the results. In this study we derived and compared several resampling-based approaches to estimate the SE. The proposed methods include parametric and non-parametric bootstrap, ordinary jackknife and delete-d-jackknife. We found that delete-d-jackknife and parametric bootstrap approaches provide good estimates of the SE. Methods to compute the sum of heritability explained and the corresponding SE are implemented in the R package SumVg, available at the author's website.

Estimating Regression Coefficients including Nuisance Baseline and its Applications (PS06)

Kenichi Kamo

Sapporo Medical University, Japan. *kamo@sapmed.ac.jp*

Abstract: Regression models can handle multiple variables all at once. In the regression models, there may be two kind of explanatory variables, one is the variables in our interest, another is one that is necessary in regression modeling but not in our interest. In such situation, we try to construct the regression models including the nuisance parameters as a baseline. It can give the estimate for the coefficients which is in our interest without some setting for the part of nuisance. Such idea appears implicitly in other regression model setting. For example, in simple regression models the coefficient for slope can be estimated without using the information for the intercept, then the intercept can be regarded as baseline. For another example, in Cox proportional hazard model, the coefficients for explanatory variables can be estimated regardless of the setting for baseline survival curve.

We apply such concept to varying coefficient models, and introduce two applications, one is to Age-Period-Cohort analysis for evaluating the risk of cancer, and another is to the growth analysis for the forest stand. In the first example, since APC model has a problem for identification, then we regard one effect (e.g. age effect) as nuisance baseline. In the second example, since the growth amount of forest is affected by several environmental factors, then we regard the geographical position and age dependent growth as nuisance baseline. In both cases, we can estimate parameters without any assumption for the baseline, i.e., nonparametrically. Besides the simplicity of modeling baseline curves, we can also describe the baseline by using the residuals after the regression coefficients have been estimated.

This work is the joint project with Tetsuji Tonda (Prefectural University of Hiroshima) and Kenichi Satoh (Hiroshima University).

Joint Inference for a GLMM Model with a NLME Covariate Model Subject to Left Censoring and Measurement Error, with Application to AIDS Studies (PS07)

Hongbin Zhang

City University of New York, United States. *hongbinzhang711@yahoo.com*

Abstract: In an HIV/AIDS study, we are often interested in the dynamics of viral load and CD4 counts over time during an antiretroviral treatment with the primary goal to understand their relationship and the interplay of a treatment option. Statistical analyses are challenging due to the fact that the viral load observations typically are left censored and measured with error. We propose a joint model for CD4 counts and viral load in which a non-linear mixed effects model (NLME) is used for the mis-measured viral load (as covariate process) and a generalized linear mixed effects model (GLMM) is used for the CD4 count (as response process) conditional on the true trajectory of the covariate process. Model parameters are jointly estimated by a Monte Carlo EM algorithm (MCEM) with Gibbs sampling. We compare the performance of our method to existing simpler methods via simulation and give an example based on a real data in which the methods do not yield the same inference on the direct (i.e., not mediated by viral load) effect of HIV antiretroviral treatment. Our simulation results suggest that our method leads to estimators with the least bias, more honest assessments of estimate uncertainty, and more accurate coverage rates.

A New Distribution to Describe Big Data (PS09)

Yuanyuan Zhang

The University of Manchester, United Kingdom. *zhangyuanyuan612@gmail.com*

Abstract: Gadepally and Kepner provided a Power law distribution to describe a big data set such as Twitter dataset and a corpus of news article provided by Reuters. Here, we reanalyse the data set and show that our new distribution provides a better fit on how big data acts. As well, our new distribution provided a better fit when testing on another new big Twitter dataset containing all the metadata associated with approximately 2 million tweets.

Nonlinear Operator Estimation with Bayes Sieve Estimator (PS10)

Masaaki Imaizumi

The University of Tokyo, Japan. *insou11@hotmail.com*

Abstract: We develop and analyze estimation method for nonlinear operator between functional spaces. The problem of the operator estimation appears in fields of functional data analysis, semi-parametric model analysis and so on. Also, estimating nonlinear operators is still a developing problem. Our estimator is based on a Bayesian nonparametric estimation method, which allows the problem of estimating nonlinear operators to be tackled by introducing a prior distribution for a spectrum of the operators. We present an analysis of consistency and convergence rate for the estimator. Given some conditions on operators, we show that the convergence rate of our estimators depends on regularities of operators and arguments. We also derive Bernstein von Mises theorem and illustrate the convergence behavior and practical viability of the estimator by simulations.

On the Spectral Distribution of Hayashi's Estimator for High Dimensional Stock Price Process (PS11)

Arnab Chakrabarti

Indian Statistical Institute, India. *arnab@isichennai.res.in*

Abstract: We consider the problem of estimation of integrated covariance for high dimensional financial stock price process. As the stock price data is nonsynchronous it is worthwhile to think about Hayashi's method of estimation (Hayashi, Yoshida 2004). The asymptotic results for small dimension is already established (Hayashi, Yoshida 2007). But in high dimensional setup when the dimension p and observation frequency grow at same rate this estimator is no longer a consistent estimator of integrated covariance. Assuming that the data is observed synchronously limiting spectral distribution of Realized Covariance matrix is established (Zeng, Li 2011). For nonsynchronous data we study of limiting behavior of the empirical spectral distribution (ESD) of high dimensional Hayashi's estimator matrix by establishing a Marcenko-Pastur type theorem.

Recent Advances in Approximate Solution for Stochastic Differential Delay Equation (PS12)

Young-Ho Kim

Changwon National University, South Korea. *yhkim@changwon.ac.kr*

Abstract: In this talk, we deal with a difference between an approximate solution and an accurate solution to the special but important class of stochastic functional differential equations which is the stochastic differential delay equations. To make the theory more understandable, we use the Caratheodory's approximate solutions to stochastic differential equations with a non-uniform Lipschitz condition and special linear growth condition.

Unified Tree-structured Non-crossing Quantile Regression Model (PS13)

Jaeoh Kim

Korea University, South Korea. *c14180@korea.ac.kr*

Abstract: Although a number of tree algorithms have been developed, very few studies have been conducted for quantile regression. Furthermore, the existing regression method can generate crossing quantiles, which is unrealistic. We propose a tree algorithm for the unified tree-structured quantile regression model with non-crossing constraints. Our tree model is constructed by recursively partitioning the data based on repeated analyses of residuals arisen from model fitting with non-crossing multiple quantile regression. In addition, our algorithm assigns the weights to each quantile function, and measures the strength of association (uncertainty coefficient) for split variable selection. The main advantage of this algorithm is that it provides a unified tree model with satisfying the non-crossing assumption of quantile regression. This result leads to an interpretable prediction model and easy data visualization. Another advantage is to dramatically reduce the selection bias of split variables and the computational cost. We investigate the performance of our algorithm with both simulated and real data compared to the previous studies (exhaustive search approach and univariate method). The simulation results show that our proposed method generates negligible bias, requires relatively less expensive computational cost, and obtains more accurate prediction. This is a joint work with HyungJun Cho, Yeonjoo Jeong, and Sungwan Bang (corresponding author).

Author Index

Name	Session	Page No.
[A]		
Aalen, Odd O.	IP48	71, 331
Abe, Hiroyasu	CP04	49, 202
Adachi, Kohei	TCP13, CP04	36, 49, 123, 201
Adavalath Puthiyaveetil, Kuttykrishnan	CP14	41, 153
Aoki, Satoshi	IP30, IP30	48, 190
Arashi, Mohammad	TCP13, TCP28	36, 53, 123, 219
[B]		
Bandyopadhyay, Soutir	IP41	58, 251
Baranowski, Rafal	TCP25	65, 295
Barigozzi, Matteo	IP40	64, 287
Belitser, Eduard	TCP01	57, 240
Bernardin, Cédric	IP13	56, 236
Berthet, Quentin	IP06	50, 205
Bhamidi, Shankar	IP14, IP44	43, 64, 163, 288
Bhattacharya, Rabi	DL07, IP13	30, 56, 82, 236
Biswas, Rahul	CP05	76, 363
Bose, Arup	DL15, IP26	63, 71, 282, 329
Breidt, Jay	DL16	59, 254
Brown, Lawrence David	IP56	48, 192
Budhiraja, Amarjit	DL07, IP13, IP15	30, 35, 56, 82, 114, 237
Buhlmann, Peter Lukas	DL08	59, 252
Buu, Anne	TCP11	61, 264
[C]		
Cai, Jingheng	TCP20, TCP32, TCP33	49, 58, 69, 199, 245, 316
Cai, Tony	IP59	68, 312
Cao, Hongyuan	IP12	35, 112
Cavaliere, Giuseppe	DL12	46, 181
Chacon, Jose E.	IP37	31, 90
Chakrabarti, Arnab	PS11	78, 376
Chan, Jennifer	IP39	74, 355
Chan, Ka Shing	CP06	58, 249
Chan, Kin Wai	CP18	76, 367
Chan, Ngai Hang	DL12	46, 180
Chan, Stephen	CP16	73, 344
Chandna, Swati	TCP31	53, 221, 222
Chang, Shu-Hui	IP48, TCP24	52, 71, 215, 332
Chang, Xiangyu	CP12	72, 339
Chang, Yu-Mei	TCP24	52, 215
Chao, Shih-Kang	TCP08	40, 148
Chatterjee, Singdhansu	IP35	40, 141
Chaudhuri, Sanjay	IP35	40, 141
Chen, Cathy W.S.	IP27	51, 208
Chen, Chen-Fang	TCP24	52, 215
Chen, Hao	TCP17	33, 101
Chen, Jinbo	TCP18	44, 171
Chen, Lin	TCP13	36, 123
Chen, Lu-Hung	IP24	51, 212
Chen, Ming-Hui	DL02, IP57	36, 42, 118, 157
Chen, Wei-Kuo	IP28	39, 137
Chen, Ying	IP36	64, 290
Chen, Ying Qing	IP48	71, 332
Chen, Ying Qing	IP48	71, 332
Chen, Yining	TCP25	65, 294
Chen, Zehua	IP03	43, 161
Chen Bauer, Cici	TCP10	45, 175, 177
Cheng, Ching-Shui	DL11	42, 155
Cheng, Dongya	CP09	66, 301
Cheng, Fengyang	CP09	66, 302
Cheng, Guang	DL01	63, 281
Cheng, Ming-Yen	IP51, IP60, PL02	1, 4, 30, 32, 38, 81, 94, 134
Cheng, Qian	TCP25	65, 295
Cheng, Tsung-Chi	DL11, IP25	42, 47, 155, 188, 189
Cheung, Mike W.-L.	TCP15	32, 96
Cheung, Siu Hung	IP04, TCP28	1, 53, 55, 219, 228
Chiou, Jeng-Min	IP24, IP25	47, 51, 188, 211
Cho, Haeran	IP40, IP40	64, 286, 287
Choi, Guebin	CP08	62, 275
Choi, Taeryon	IP19, IP54	44, 47, 170, 184
Christmann, Andreas	IP34	31, 88
Chu, Chi Wing George	TCP23	41, 149
Chu, Jeffrey	CP08	62, 277
Chu, Weijuan	TCP08	40, 146
Chung, Ray S W	CP06	58, 247
Clinet, Simon	TCP02	65, 292
Cox, Dennis	IP19	47, 185
Cuesta-Albertos, Juan Antonio	IP34	31, 88
[D]		
Dai, Wenlin	CP01	33, 104
Daniels, Michael	IP52	60, 262
Davies, Tilman M.	IP37	31, 90
Davis, Richard A.	IP16, IP17	39, 74, 135, 352
Davison, Anthony	IP16	39, 135
Dayaratna, Kevin	CP07	75, 361
de Carvalho, Miguel	IP16	39, 135
Deng, KE	IP49	67, 309
Dhar, Subhra Sankar	IP34	31, 88
Diaconis, Persi	PL01	4, 30, 80
Dini, Aulia	CP13	76, 364
Dou, Xiaoling	TCP20	49, 199
Du, Chao	IP04	55, 228
Du, Lilun	CP04	49, 202
[E]		
Effendie, Adhitya Ronnie	CP13	76, 365
Effraimidis, Georgios	CP19	77, 370
Eguchi, Shoichi	CP02	37, 129
Elliott, Michael	IP35	40, 142
Engelke, Sebastian	IP16	39, 135
[F]		
Fan, Jianqing	DL17, IP60	34, 38, 108, 134
Fan, Wai Tong	DL05	67, 305
Fan, Xiaodan	IP49, CP01, TCP26	1, 33, 61, 67, 103, 267, 308
Fan, Yingying	IP06	50, 205
Fang, Xiao	IP44	43, 163

Name	Session	Page No.
Farinetto, Christian	CP14	41, 154
Feng, Xiang-Nan	TCP33	69, 316
Feng, Yang	IP05, IP51	32, 46, 94, 183
Fiecas, Mark	IP40	64, 286
Fine, Jason	IP07	34, 110
Franco Saldana, Diego	TCP31	53, 222
Fujisawa, Hironori	CP04, CP05	49, 76, 201, 362
Funaki, Tadahisa	IP13	56, 236
Fung, Wing Kam	IP43	60, 261
[G]		
Gan, Fah Fatt	CP11, CP12	70, 72, 324, 339
Ganguly, Arnab	IP15	35, 114
Gao, Fuchang (Frank)	TCP29	61, 272
Gao, Xiaoli	IP34	31, 89
Geenens, Gery	IP37	31, 91
Genton, Marc G.	IP33, IP41	58, 72, 250, 337
Ghosh, Malay	IP02, IP35	40, 42, 141, 160
Ghoshal, Subhashis	TCP01	57, 240, 241
Giraitis, Liudas	IP17, IP27	51, 74, 207, 352
Gou, Jianwei	TCP27	52, 217
Gu, Lijie	CP15	53, 223
Gu, Ming Gao	TCP05	69, 321
Guo, Guangbao	CP03	45, 179
Guo, Jing	TCP35	66, 298
Guo, Shaojun	IP59, TCP25	65, 68, 294, 312, 313
Guo, Xu	TCP19	45, 173
[H]		
Hall, Daniel	IP18	43, 166
Hallin, Marc	DL12	46, 180
Han, Zhongcheng	CP01	33, 103
Hannig, Jan	DL03	55, 226
Hao, Hongxia	CP14	41, 154
Hara, Hisayuki	IP30	48, 190
Hashiguchi, Hiroki	TCP22	49, 197
Hayter, Anthony	TCP28	53, 219
Hazelton, Martin	IP37	31, 90, 91
He, Kevin (Zhi)	IP42, TCP05	56, 69, 232, 321
He, Xuming	IP43	1, 60, 260
Hernita, Tudzla	CP09	66, 301
Heuchenne, Cedric	IP42	56, 233
Hirose, Kei	TCP03	40, 143
Hirose, Yoshihiro	TCP06	41, 151
Hirukawa, Masayuki	IP31	52, 214
Honda, Toshio	IP12, IP31	35, 52, 113, 213
Hou, Lin	IP49	67, 309
Hsieh, Jin-Jian	CP17	73, 348
Hsu, Hsien-Yuan	TCP15	32, 96
Hu, Feifang	IP09, IP29	51, 63, 210, 284
Hu, Guanyu	CP17	73, 347
Hu, Jianhua	IP07	34, 110
Hu, Jie	TCP01, TCP34	57, 65, 240, 296
Hu, Qilin	TCP25	65, 294

Name	Session	Page No.
Hu, Yaozhong	IP58	75, 358
Hu, Yu-Pin	CP16, CP18	73, 76, 345, 366
Hu, Zonghui	IP52	60, 262, 263
Huang, Alan	IP39	74, 354
Huang, Chiung-Yu	IP52	60, 262
Huang, Danyang	DL10	34, 106
Huang, Jian	IP54	44, 169
Huang, Li-Shan	IP10	56, 234
Huang, Ruifen	TCP35	66, 298
Huang, Tao	IP21	32, 93
Huang, Yangxin	CP06	58, 248
Huber-Carol, Catherine Louise	IP48	71, 331
Huckemann, Stephan	DL07	30, 82
Huser, Raphael	IP16	39, 136
Hwang, Hsien-Kuei	IP32	39, 139
Hwang, Wei-Ting	TCP09	68, 314
[I]		
Iacus, Stefano	DL14	74, 351
Ikemoto, Hiroki	TCP13	36, 124
Iki, Kiyotaka	CP10	73, 343
Imaizumi, Masaaki	PS10	78, 376
Ing, Ching-Kang	DL06	50, 203
Ishii, Aki	CP08	62, 275
Isopi, Marco	DL15	71, 329
Iwasa, Manabu	TCP13	36, 124
[J]		
Jang, Woncheol	DL09	1, 30, 84
Jasra, Ajay	TCP03	40, 144
Ji, Hongkai	IP42, IP49	56, 67, 233, 308
Jiang, Bei	TCP16	33, 99
Jiang, Ci-Ren	IP24	51, 211
Jiang, Han	TCP30, CP03	45, 62, 178, 273
Jiang, Hangjin	CP07	75, 360
Jiang, Hui	TCP11	61, 264
Jiang, Peiyong	TCP26	61, 267
Jiang, Xuejun	TCP33, TCP34	65, 69, 296, 317
Jin, Zhezhen	IP10, IP11	56, 59, 234, 235, 256
Jing, Bingyi	IP53, IP60	38, 72, 134, 335
Johnstone, Iain	PL02	4, 30, 81
Jung, Sungkyu	IP38, TCP17	33, 44, 101, 167
Jung, Yoonsuh	CP02	37, 128
[K]		
Kakizawa, Yoshihide	IP31	52, 213
Kamatani, Kengo	TCP03	40, 143, 145
Kamo, Kenichi	PS06	78, 374
Kao, Ming-Hung	DL11	42, 155
Katayama, Shota	CP03	45, 178
Kawano, Shuichi	CP05	76, 362
Ke, Zheng Tracy	IP05	46, 182
Kim, ByungWon	TCP17	33, 101
Kim, Jaeoh	PS13	78, 377
Kim, Kunwoo	IP15	35, 115

Name	Session	Page No.
Kim, Minjung	TCP14	37, 126
Kim, Mi-Ok	IP18, IP43	43, 60, 165, 260
Kim, Soeun	IP18	43, 166
Kim, YangJin	CP17, CP19	73, 77, 349, 369
Kim, Yongdai	TCP01	57, 241
Kim, Young-Ho	PS12	78, 377
Kobayashi, Kei	TCP31	53, 221
Koch, Inge	IP38	44, 167, 168
Koike, Yuta	TCP02	65, 292
Kolaczyk, Eric	IP14	64, 288
Komaki, Fumiyasu	TCP07	69, 319
Kong, Linglong	TCP16	33, 99
Kong, Xinbing	IP53	72, 335
Koo, Ja-Yong	DL09	30, 84
Kosorok, Michael	DL01	63, 280
Kou, Samuel	IP04	55, 228, 229
Kumar, Rohini	IP15	35, 115
Kundu, Subrata	IP33	72, 338
Kuriki, Satoshi	IP30, IP32	39, 48, 139, 190
Kuroda, Masahiro	TCP22	49, 196
Kurtz, Thomas G.	DL05	67, 304
Kuwabara, Tatsuya	TCP22	49, 196
Kwok, Oi-Man	TCP14, TCP15	32, 37, 96, 126
Kyung, Minjung	IP19	47, 185
[L]		
Lai, Hok Chio	TCP15	32, 97
Lai, Tze Leung	DL06	50, 203
Lan, Xiaohong	IP45	67, 306
Landis, J. Richard	IP42	56, 232
Lee, Eun Ryung	IP40	64, 286
Lee, J. Jack	IP09	63, 284
Lee, Sangyeol	IP17	74, 353
Lee, Stephen	TCP30	62, 273
Lee, Taewook	TCP32	58, 245
Lee, Yoonkyung	CP03, CP05	45, 76, 178, 362
Lee, Young K.	IP20, IP40	35, 64, 116, 286
Lee, Yuan-Hsuan	TCP15	32, 97
Li, Cuixia	IP53	72, 336
Li, Daoji	IP06, IP11	50, 59, 205, 257
Li, Fang	IP47, CP02	37, 68, 128, 310
Li, Guodong	IP27	51, 207
Li, Han	TCP26, TCP34	61, 65, 267, 296
Li, Jialiang	IP11, IP12	35, 59, 112, 256
Li, Ken W	CP13	76, 365
Li, Ming	TCP11	61, 264
Li, Quefeng	TCP04	57, 242
Li, Rui	CP08	62, 276
Li, Runze	IP01, IP50, TCP11	31, 61, 75, 86, 264, 356
Li, Shanshan	TCP09	68, 314
Li, Wai Keung	DL12	46, 180
Li, Wai Ming	CP06	58, 248
Li, Yan	TCP18	44, 171

Name	Session	Page No.
Li, Yanming	TCP05	69, 321
Li, Yehua	IP12, IP18	35, 43, 113, 165
Li, Yi	IP42, IP50, TCP05	56, 69, 75, 232, 321, 356
Li, Yingying	IP05	46, 109
Li, Zhaohai	TCP18	44, 171
Lian, Heng	IP03	43, 162
Liang, Faming	IP54	44, 169
Lim, Chae Young	IP41	58, 250
Lim, Johan	DL06	50, 204
Lin, Chien-Tai	TCP12	36, 120
Lin, Huazhen	IP46	71, 334
Lin, Lizhen	DL07	30, 83
Lin, Shili	DL04	70, 327
Lin, Yuanyuan	IP21	32, 92
Ling, Man Ho	TCP12	36, 120
Lipsitz, Stuart	IP43	60, 260
Liu, Anna	TCP30	62, 273
Liu, Jin	TCP34	65, 297
Liu, Mengling	DL03, IP43	55, 60, 226, 260
Liu, Min-Qian	IP23	47, 186
Liu, Pengfei	TCP33	69, 317
Liu, Regina Y.	DL03, DL08	55, 59, 226, 227, 252
Liu, Weidong	IP28	39, 137
Liu, Weiwei	CP01	33, 104
Liu, Xuhua	TCP21	48, 194
Liu, Yan	TCP34, TCP35, CP06	58, 65, 66, 247, 297, 298
Liu, Yingan	TCP27	52, 217
Liu, Yong	TCP29	61, 270
Liu, Zhi	IP53	72, 336
Liu, Zhongqiang	IP09	63, 285
Low, Mark	IP59	68, 312
Lu, Tong-Yu	TCP28	53, 220
Lu, Wenbin	DL01, IP46	63, 71, 280, 333
Lumley, Thomas	IP35	40, 142
Luo, Sheng	CP19	77, 370
Luo, Wen	TCP15	32, 98
[M]		
Ma, Ping	IP49	67, 308
Ma, Shuangge	IP55	57, 239
Ma, Yingying	CP11	70, 325
Maesono, Yoshihiko	IP31	52, 213
Maiti, Taps	IP41	58, 250
Majumdar, Anandamayee	IP33	72, 338
Masuda, Hiroki	DL14, TCP02, TCP03	40, 65, 74, 144, 292, 350
Matsuda, Takeru	CP16	73, 344
Matsui, Hidetoshi	CP01	33, 105
Matsuzoe, Hiroshi	TCP06	41, 151
Meng, Helen	IP29	51, 209
Mengersen, Kerrie	DL02	42, 157
Minin, Vladimir	DL04	70, 328
Miyazaki, Mitsuhiro	TCP13	36, 125
Mori, Yuichi	TCP22	49, 196
Moriyama, Taku	IP31	52, 214

Name	Session	Page No.
Mueller, Samuel	IP39	74, 354
Mukherjee, Ayon	CP18	76, 368
Mukherjee, Gourab	IP11, IP36, TCP07	59, 64, 69, 256, 290, 319
Mukherjee, Partha Sarathi	CP11	70, 324
[N]		
Nagatsuka, Hideki	TCP12	36, 121
Naito, Kanta	IP38	44, 168
Nakagawa, Shigekazu	TCP22	49, 197
Nandy, Rajesh Ranjan	IP33	72, 337
Neeman, Joe	IP44	43, 164
Ng, Chi Tim	TCP21, TCP32	48, 58, 194, 245
Ng, Tony Hon Keung	TCP12	36, 120
Nicholls, Geoff	DL02	42, 157
Niu, Cui Zhen	TCP19	45, 173
Noh, Hohsuk	IP20	35, 116
Nott, David John	IP19	47, 184
[O]		
Ogawa, Mitsunori	CP15	53, 224
Ogihara, Teppei	TCP02, TCP03	40, 65, 143, 293
Ong, Meng Hwee Victor	DL16	59, 255
Ota, Yasushi	TCP22	49, 198
Ouyang, Shun-Xiang	TCP29	61, 270
Ouyang, Yunbo	TCP07	69, 319
[P]		
Pan, Deng	TCP20	49, 199
Pan, Guangming	IP26	63, 282
Pan, Junhao	TCP27, TCP32, TCP33	52, 58, 69, 217, 246, 316
Pan, Maolin	TCP19, TCP33	45, 69, 173, 318
Pang, Zhen	IP39	74, 355
Paolella, Marc	IP27	51, 208
Park, Byeong	IP56, IP60	38, 48, 134, 192
Pati, Debdeep	IP02	42, 159
Patilea, Valentin	IP38, IP56	44, 48, 167, 193
Paul, Debashis	IP26, IP33	63, 72, 282, 337
Penev, Spiridon Ivanov	IP31, IP37	31, 52, 90, 213
Peng, Jie	IP43, TCP17	33, 60, 101, 261
Peng, Limin	IP46	71, 334
Peng, Yingwei	IP03	43, 162
Peng, Yong	TCP35	66, 299
Petrovic, Sonja	IP30	48, 191
Pham, Tung	CP01, CP02	33, 37, 103, 128
Phoa, Frederick Kin Hing	DL11	42, 156
[Q]		
Qian, Cheng	TCP25	65, 295
Qiao, Xinghao	IP59, TCP25	65, 68, 294, 313
Qiu, Yumou	TCP05	69, 322
Qu, Annie Peiyong	DL13	38, 132
Qu, Leming	CP17, CP19	73, 77, 347, 369

Name	Session	Page No.
[R]		
Ratcliffe, Sarah J.	IP18	43, 165
Ren, Zhao	IP06	50, 206
Richards, Donald	IP30, IP32	39, 48, 139, 191
Roeckner, Michael	IP15	35, 114
Roellin, Adrian	IP14, IP44	43, 64, 163, 289
Rosenberger, William	IP09	63, 284
Ross, Nathan	IP44	43, 164
Rothe, Christoph	IP20	35, 117
Roy, Rahul	DL15	71, 329
Ryu, Ehri	TCP14	37, 126
[S]		
Sadeghi, Kayvan	TCP31	53, 221
Saigusa, Yusuke	CP10	73, 342
Sajadi, Farkhondeh Alsadat	DL15	71, 330
Sakata, Toshio	TCP13	36, 123
Sang, Hailin	CP02	37, 129
Sang, Huiyan	TCP16	33, 100
Satoh, Kenichi	PS04	78, 373
Scealy, Janice Lea	IP39	74, 354
Schienle, Melanie	IP20	35, 117
Sei, Tomonari	TCP06	41, 151
Sen, Sanchayan	IP14	64, 288
Sengupta, Debasis	IP33	72, 337
Sethuraman, Sunder	DL05, IP13	56, 67, 236, 237, 304
Shah, Rajen Dinesh	DL08	59, 253
Shao, Qi-Man	IP28, IP29, PL01	39, 51, 80, 137, 209
She, Yiyuan	IP01, IP02	31, 42, 87, 159
Shen, Guangjun	TCP08	40, 147
Shen, Haipeng	IP24, IP56	48, 51, 192, 212
Shen, Jing	TCP27	52, 217
Shen, Linghao	CP16	73, 345
Shen, Weining	TCP01	57, 240
Shi, Jianxin	TCP18	44, 172
Shimizu, Yusuke	TCP02	65, 293
Shojaie, Ali	DL08	59, 252
Siegmund, David	IP04	55, 229
Singh, Vikas	IP08	55, 230
Sinha, Debajyoti	IP02	42, 159
Sit, Tony	TCP10, TCP23	41, 45, 149, 175
So, Hon Yiu	TCP12	36, 121
So, Hon-Cheong	PS05	78, 374
Song, Jian	IP58	75, 358
Song, Peter XK	IP03, IP07	34, 43, 111, 161
Song, Rui	IP11, IP46	59, 71, 257, 333
Song, Xin Yuan	DL13, IP54, TCP20	44, 49, 132, 169, 199
Stoev, Stilian	IP16	39, 136
Su, Deng-Huang	TCP24	52, 216
Su, Jingyong	TCP17	33, 102
Su, Zhonggen	IP05, TCP29	46, 61, 183, 270
Sun, Yanqing	IP10	56, 234
Sun, Ying	IP41	58, 251
Suprihatin, Bambang	CP16	73, 346

Name	Session	Page No.
[T]		
Tahata, Kouji	CP08, CP10	62, 73, 275, 343
Takemura, Akimichi	IP32	39, 139
Tamatani, Mitsuru	CP10	73, 342
Tan, Zhiqiang	IP52	60, 262
Tanaka, Kentaro	TCP06	41, 152
Tanaka, Ushio	TCP06	41, 152
Tang, Yu	IP23	47, 186
Tarr, Garth	IP34	31, 89
Thompson, Elizabeth	DL04	70, 327
Tian, Guo-Liang	TCP10	45, 176
Ting Lee, Mei-Ling	IP48	71, 331
Tonda, Tetsuji	PS02	78, 372
Tong, Xin	IP06, IP59	50, 68, 206, 312
Truong, Chau Buu	CP17	73, 347
Truquet, Lionel	IP17	74, 353
Tsai, Chen-An	IP25	47, 189
Tsai, Shin-Fu	CP15	53, 224
Tseng, Yi-Kuan	IP24, IP25	47, 51, 188, 211
Tsukahara, Hideatsu	CP19	77, 369
[U]		
Uchida, Masayuki	DL14	74, 351
Uehara, Yuma	TCP08	40, 147
Uhler, Caroline	IP32	39, 140
[V]		
Van Keilegom, Ingrid	IP50	75, 357
[W]		
Wagenmakers, Eric-Jan	DL02	42, 158
Wand, Matt	DL16, IP60	38, 59, 134, 254
Wang, Bo	IP19	47, 184
Wang, Chaojie	CP11	70, 326
Wang, Chunjie	TCP20	49, 200
Wang, Dewei	TCP04	57, 242
Wang, Dong (NUS)	TCP29	61, 271
Wang, Feng-Yu	IP58	75, 359
Wang, Guochang	TCP20	49, 200
Wang, Hansheng	DL10	34, 106
Wang, Jing	IP47	68, 310
Wang, Jiun-Yi	TCP24	52, 216
Wang, Lan	IP01, IP50	31, 75, 87, 356
Wang, Maggie	CP07	75, 360
Wang, Man	TCP32	58, 246
Wang, Qihua	IP52	60, 263
Wang, Qinwen	IP26	63, 283
Wang, Rachel	TCP17	33, 102
Wang, Sijian	TCP16	33, 99, 100
Wang, Wensheng	IP45	67, 306
Wang, Xiao	IP07	34, 111
Wang, Xiaofei	CP17	73, 348
Wang, Xiaoning	TCP21	48, 194
Wang, Xueqin	IP08, IP54	44, 55, 170, 230
Wang, Xuexia	CP11, CP12	70, 72, 324, 341
Wang, Yuan	TCP04	57, 243
Wang, Yuedong	IP42	56, 232
Watanabe(Chang), Genso(Yuan-Tsung)	CP11	70, 325
Wei, Yingying	IP05, IP55, TCP09	46, 57, 68, 182, 238, 315
Weinstein, Asaf	TCP07	69, 320
Welsh, Alan	DL16, IP60	38, 59, 134, 254
Wilson, Susan	DL04, IP07, IP60	34, 38, 70, 110, 134, 327
Wolters, Mark	CP18	76, 366
Won, Joong-Ho	DL09	30, 85
Wong, C.K.	IP29	51, 209
Wong, Hoi Ying	DL10	34, 106
Wong, Kin Yau	IP46	71, 333
Wong, Samuel W.K.	IP04	55, 228
Wu, Changbao	TCP10	45, 176
Wu, Dongsheng	IP45, TCP08	40, 67, 146, 307
Wu, Hao	IP55	57, 238
Wu, Jiun-Yu	TCP14	37, 127
Wu, Rongling	TCP11	61, 265
Wu, Yilei	TCP30	62, 274
Wu, Zhengxiao	CP14, CP15	41, 53, 153, 223
[X]		
Xia, Lucy	IP51	32, 95
Xia, Yemao	TCP27	52, 218
Xia, Yin	IP57	36, 119
Xiang, Liming	IP12	35, 112
Xiao, Yimin	IP45	67, 306, 307
Xie, Min-ge	DL03	55, 226
Xiong, Jie	DL05	67, 304
Xiong, Shifeng	IP23	47, 187
Xu, Gongjun	TCP23	41, 149
Xu, Lihu	IP58	75, 359
Xu, Peirong	TCP05	69, 322
Xu, Wangli	IP22	60, 258
Xue, Lan	IP47	68, 310, 311
Xue, Yuan	IP22	60, 259
[Y]		
Yam, Phillip Sheung Chi	CP04, CP09, IP50	49, 66, 75, 201, 301, 357
Yamamoto, Kouji	CP18	76, 366
Yan, Jigao	CP09	66, 303
Yan, Liang	TCP21	48, 195
Yang, Can	IP55	57, 238
Yang, Dan	IP56, IP57	36, 48, 118, 193
Yang, Guangren	TCP11	61, 265
Yang, James	TCP11	61, 266
Yang, Jian-Feng	IP23	47, 186
Yang, Lijian	IP47	68, 311
Yang, Ping	TCP28	53, 220
Yang, Qiang	IP29	51, 210
Yang, Yi-Hsin	IP25	47, 188
Yano, Keisuke	CP06	58, 247
Yao, Jianfeng	IP26	63, 282, 283
Yao, Zhigang	IP36	64, 291
Yau, Chun Yip	CP14, CP16	41, 73, 153, 344

Name	Session	Page No.
Ye, Zhisheng	TCP12	36, 122
Yin, Xiangrong	IP38	44, 167
Yoon, Myeongsun	TCP14	37, 127
Yoshida, Nakahiro	DL14	74, 350
Yu, Donghyeon	CP12	72, 339
Yu, Kai	TCP18	44, 172
Yu, Kyusang	IP20	35, 116
Yu, Philip L.H.	IP27	51, 207
Yu, Sheng	IP49	67, 308
Yu, Tao	IP36	64, 291
Yu, Tianwei	CP07, CP12	72, 75, 340, 360
Yu, Weichuan	TCP26	61, 268
Yu, Wen	TCP23	41, 150
Yu, Yi	TCP23, TCP31	41, 53, 150, 221
Yu, Zhou	IP22	60, 258
Yuan, Ming	IP06	50, 205
Yuan, Shun	TCP35	66, 299
[Z]		
Zangeneh, Sahar	TCP10	45, 175
Zhang, Anderson Ye	TCP07	69, 320
Zhang, Baqun	TCP05	69, 323
Zhang, Chongqi	IP23	47, 186
Zhang, Chunming	IP08	55, 230
Zhang, Cun-Hui	DL17, IP01	31, 34, 86, 108
Zhang, Deng	TCP29	61, 271
Zhang, Hao	IP08	55, 231
Zhang, Heping	IP03	43, 161
Zhang, Hong	TCP18	44, 171
Zhang, Hongbin	PS07	78, 375
Zhang, Jiangwen	TCP26	61, 268
Zhang, Jin-Ting	DL06, IP36	50, 64, 203, 290
Zhang, Kai	TCP30	62, 274
Zhang, Kaimeng	PS03	78, 373
Zhang, Li-Xin	IP09	63, 285
Zhang, Lingsong	IP57, TCP16	33, 36, 99, 119
Zhang, Wenyang	IP10, IP11, IP17	56, 59, 74, 234, 256, 352
Zhang, Xin	IP02	42, 160
Zhang, Xinsheng	IP53	72, 335
Zhang, Yongli	IP47	68, 310
Zhang, Yuanyuan	PS09	78, 375
Zhang, Zheng	IP22, CP10, CP13	60, 73, 76, 259, 342, 364
Zhang, Zhengjun	IP08, IP45	55, 67, 230, 231, 306,
Zhang, Zhuosong	IP28	39, 138
Zhao, Hongyu	IP55	57, 239
Zhao, Junlong	TCP21	48, 194
Zhao, li	TCP21	48, 195
Zhao, Linda	IP57	36, 118
Zhao, Xin	TCP35	66, 298
Zhao, Xingqiu	IP10, IP48	56, 71, 235, 331
Zheng, Hui	TCP35	66, 300
Zheng, Qi	TCP04	57, 243
Zheng, Tian	IP51	32, 94

Name	Session	Page No.
Zheng, Xiaoqi	TCP26	61, 269
Zheng, Xinghua	DL17	34, 109
Zheng, Zemin	TCP04	57, 244
Zhou, Harrison	DL17	34, 108
Zhou, Jing	CP13	76, 364
Zhou, Jingke	TCP19	45, 174
Zhou, Wenxin	IP28	39, 138
Zhu, Degang	TCP27	52, 218
Zhu, Hong	TCP09	68, 315
Zhu, Hongtu	DL09	30, 84
Zhu, Ji	IP51	32, 95
Zhu, Liping	DL13, IP21, IP22	32, 38, 60, 92, 133, 258
Zhu, Lixing	DL13	38, 132
Zhu, Qianqian	IP27	51, 207
Zhu, Ruoqing	DL01, TCP09	63, 68, 280, 314
Zhu, Xuehu	TCP19	45, 174
Zhu, Xuening	DL10	34, 107
Zhu, Zhongyi	IP21, IP22	32, 60, 92, 258
Zou, Hui	IP01	31, 86

Memo

Memo

Memo

Organized by:

Institute of Mathematical Statistics

Hosted by:

Department of Statistics,
The Chinese University of Hong Kong