12th ICSA International Conference

Hong Kong, China

July 7- 9, 2023

International Chinese Statistical Association

泛華統計協會

International Chinese Statistical Association

12th International Conference
2023

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

July 7 - 9, 2023

The Chinese University of Hong Kong

Hong Kong, China

Organized by

International Chinese Statistical Association

Contents

12th ICSA International Conference
July 7 - 9, 2023
Organized at The Chinese University of Hong Kong, Hong Kong

Welcome to the 12th International Chinese Statistical Association (ICSA) International Conference!

As you know, the 12th ICSA International Conference was originally planned to be held in December 2022 and postponed to the current dates due to the COVID-19 pandemic. We are so glad that we can meet in person without any restrictions now. For this to happen, the local organizing committee members, faculty, students and staff at the Department of Statistics at The Chinese University of Hong Kong, have spent many hours on many aspects including conference rooms and lunches as well as the banquet. It is for sure that the conference would not be possible without their hard work.

The organizing committees have been working diligently to put together a comprehensive scientific program and other activities to provide ample opportunities for discussions and exchanges of novel ideas in advancing statistics research and applications. The conference program contains over 100 scientific sessions, two named lectures, one keynote lecture and two special invited lectures. The two named lectures are Peter Hall lecture given by *Dr. Song Xi Chen (Peking University, China)* and Pao-Lu Hsu lecture given by *Dr. Ji Zhu (University of Michigan, USA)*. The keynote lecture will be given by *Dr. Qi-Man Shao (Southern University of Science and Technology, China)* and the two special invited lecturers are *Dr. Mei-Ling Lee (University of Maryland, USA)* and *Dr. Gang Li (UCLA, USA)*. The conference highlights methodological and applied contributions of statistics, data science, mathematics, and computer sciences. It brings together the statistical and data science communities as well as scientists from related fields to present, discuss and disseminate research and best practice.

With your full support, this conference attracts over 400 statisticians and data scientists working in academia, government, and industry from all over the world. We hope that the conference offers you great opportunities for learning, networking and recruiting, and that you will receive inspiration from the presented research ideas and develop new ones. We believe that this conference will be a memorable, interesting and enjoyable experience for all of us.

In addition to the conference, we hope that you will find time to enjoy the city of Hong Kong which is famous for many attractions, including business, international trade, entertainment, culture, food, media, fashion, science, technology, education, medicine, and research. Some of you may be here before and some may be the first timer. We are sure that all of you will find something to enjoy after a long COVID-19 pandemic lockdown or staying-home.

Thank you for coming to the 12th ICSA International Conference!

Xinyuan Song and (Tony) Jianguo Sun, on behalf of the 12th ICSA 2022 International Conference Executive and Organizing Committees

Executive Committee:
- Xinyuan Song (Co-Chair) (The Chinese University of Hong Kong)
- (Tony) Jianguo Sun (Co-Chair) (University of Missouri)
- Zhenzhen Jin (Columbia University)
- Gang LI (UCLA),
- Jun Zhao (Antengene)

Scientific Program Committee:
- (Tony) Jianguo Sun (Co-Chair) (University of Missouri)
- Xingqiu Zhao (Co-Chair) (The Hong Kong Polytechnic University)
- Cai, Jianwen (University of North Carolina-Chapel Hill)
- Cao, Hongyuan (Florida State University)
- Chang, Shu-Hui (張淑惠) (National Taiwan University)
- Chen, (Ding-Geng) Din (University of North Carolina-Chapel Hill)
- Chen, Yi-Hau (Institute of Statistical Science, Academia Sinica)
- Fan, Jianqing (Princeton University)
- Fang, Hongbin (Georgetown University)
- Hu, Joan (Simon Fraser University)
- Huang, Xuelin (The University of Texas MD Anderson Cancer Center)
- Kim, Jaehee (Duksung Women's University)
- Kim, Yangjin (Sookmyung Women University)
- Lee, Mei-Ling (University of Maryland)
- Li, Gang-Eisai
- Li, Gang (University of California, Los Angeles)
- Li, Jialiang (National University of Singapore)
- Li, Runze (Pennsylvania State University)
- Li, Yi (University of Michigan)
- Lin, Danyu (University of North Carolina at Chapel Hill)
- Lin, Huazhen (Southwestern University of Finance and Economics)
- Lin, Yuanyuan (The Chinese University of Hong Kong)
- Ning, Jing (The University of Texas MD Anderson Cancer Center)
- Pan, Jianxin (The University of Manchester)
- Shen, Wei-Lily (Eli Lilly and Company)
- Song, X.K.(Peter) (University of Michigan)
- Song, Xinyuan (The Chinese University of Hong Kong)
- Sun, Ryan (The University of Texas MD Anderson Cancer Center)
- Sun, Yanqing (University North Carolina, Charlotte)
- Tang, Niansheng (Yunnan University)
- Wang, Jane-Ling (University of California, Davis)
- Wang, Lu (University of Michigan)
- Wang, Mei-Cheng (Johns Hopkins University)
- Wang, Tao (Shanghai Jiao Tong University)
- Yang, Can (HK University of Science and Technology)

- Yin, Guosheng (University of Hong Kong)
- Zeng, Donglin (University of North Carolina)
- Zhang, Heping (Yale University)
- Zhang, Emma Jingfei (University of Miami)
- Zhang, Ying (University of Nebraska)
- Zhao, Hongyu (Yale University)
- Zhao, Yichuan (Georgia State University)

Junior Research Award Committee:
- Hongbin Fang (Chair), Georgetown University
- Chun-ling Liu, The Hong Kong Polytechnic University
- Chunjie Wang, Changchun University of Technology
- Bin Zhang, Cincinnati Children's Hospital and Medical Center

Program Book Committee:
- Dayu Sun (Chair), Indiana University/Emory University
- Yuanyuan Guo, Biogen
- Mingkai Wang, The Hong Kong Polytechnic University

IT and Website Committee:
- Yuanyuan Guo (Co-Chair), Biogen
- Chengsheng Jiang (Co-Chair), Flatiron Health
- Julian Wong, The Chinese University of Hong Kong
- Dayu Sun, Indiana University/Emory University
- BaiHan Zhao, The Hong Kong Polytechnic University

Local Organization Committee:
Accommodation (hotel arrangements, enquires related to accommodation):
- Prof. Wong Hoi Ying (Main Coordinator)
- Dr. Wong Tat Wing
- Yanny Ng

Budget & Accounting(control and account for all financial transactions, manage income and expenditures):
- Prof. Song Xinyuan (Main Coordinator)
- Yanny Ng

Catering & Banquet(arrangement of lunch and refreshments, arrangement of Banquet):
- Prof. Lin Yuanyuan (Main Coordinator
- Dr. Liu Kin Yat
- Dr. Ouyang Ming

Enquires and help centre(manage all incoming enquires and oversee the completion of each inquiry case):
- Prof. Dai Ben (Main Coordinator)
- Prof. Fang Xiao

- Dr. Leung Sze Him Isaac
- Prof. Phillip Yam
- Prof. Yau Chun Yip
- Prof. Zhu Huichen

External promotion (design and dissemination of promotional materials for the conference, continually updating the information on the conference and informing registered participants and potential participants, press release):

- Prof. Lin Zhixiang (Main Coordinator)
- Dr. Chan Chun Man
- Prof. Fan Xiaodan
- Prof. Tony Sit
- Prof. Wei Yingying
- Dr. John Wright
- Prof. Yau Chun Yip

Manpower and on-site support (recruitment of student helpers, logistic arrangement, including transportation):

- Prof. Wang Junhui (Main Coordinator)
- Dr. Chan Chun Man
- Prof. Chan Ping Shing Ben
- Dr. Cheung King Chau
- Prof. Dai Ben
- Dr. Ho Kwok Wah
- Prof. Lin Zhixiang
- Dr. Ouyang Ming
- Prof. Song Xinyuan

Souvenirs (souvenir for keynote speakers, design and production of the souvenir for participants):

- Prof. Chan Kin Wai (Main Coordinator)
- Prof. Lin Yuanyuan

Symposium Website:

- Prof. Fan Xiaodan (Main Coordinator)
- Prof. Chan Kin Wai
- Prof. Tony Sit
- Prof. Song Xinyuan
- Julian Wong

Wi-Fi Information

Wi-Fi SSID :   CUguest
Please select: Conference Guests (@conference.cuhk.edu.hk)
Login User ID :   icsa2023
Login Password : cuHK#%23

Conference Venue - Henry Cheng International Conference Centre

Address: Cheng Yu Tung Building, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

University Station

CUHK Medical Centre

MTR Exit B

Bus Terminus

Henry Cheng International Conference Centre, Cheng Yu Tung Building

Cheng Yu Tung Bldg.

To Hung Hom

Hyatt Regency Hong Kong, Sha Tin (Hotel Facilities of the Teaching Hotel of The Chinese University of Hong Kong)

CUHK Campus Map

CUHK Campus Map
https://www.cuhk.edu.hk/english/campus/cuhk-campus-map.html
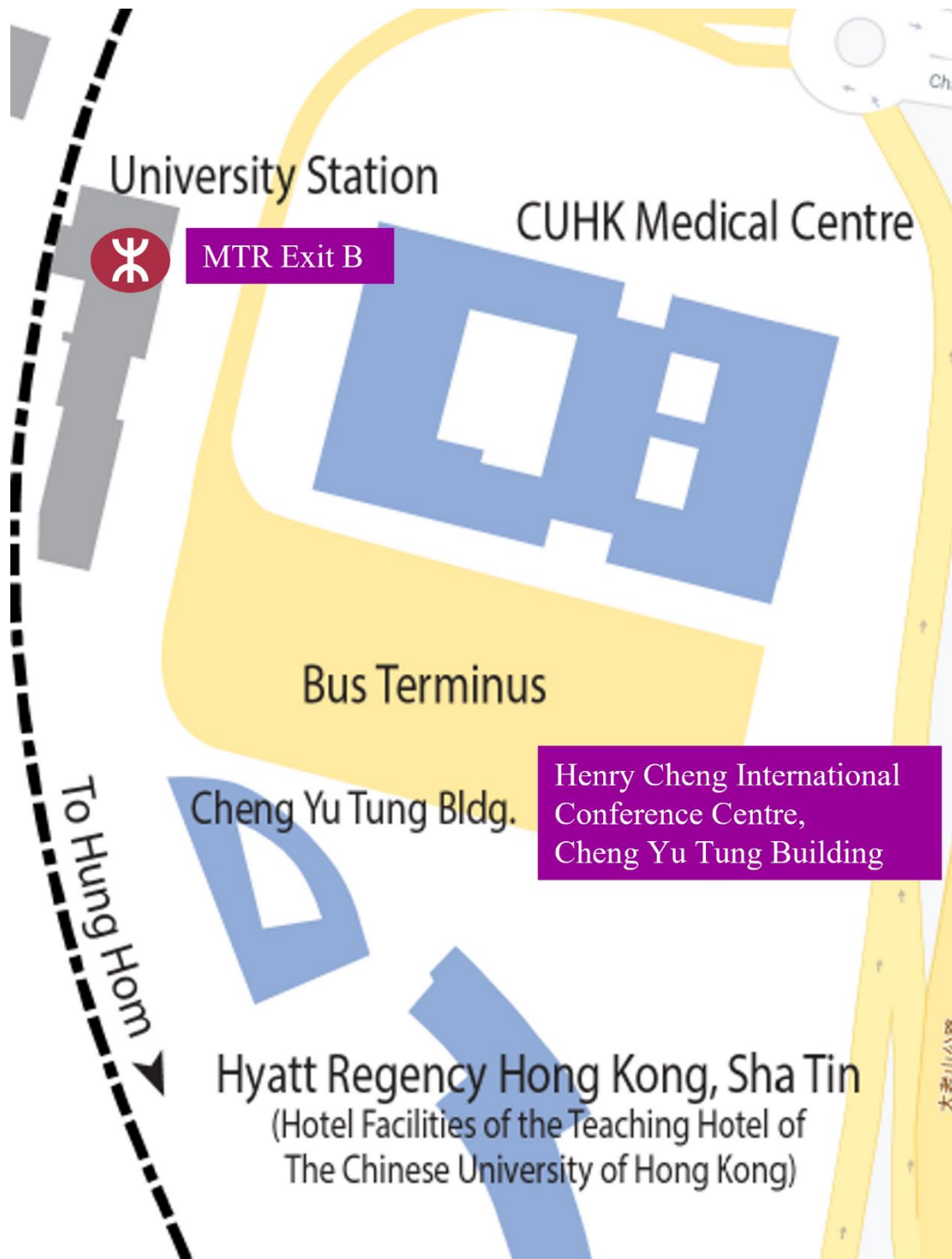
MTR System Map

1 University Station
i) The Chinese University of Hong Kong
ii) Hyatt Regency Hong Kong, Shatin

2 Shek Mun Station
i) Banquet Venue - ClubONE Riviera, Shatin
ii) Alva Hotel By Royal
iii) Courtyard by Marriott Hong Kong Sha Tin

3 Sha Tin Station
i) Royal Park Hotel
ii) Regal Riverside Hotel

MTR System Map

https://www.mtr.com.hk/en/customer/services/system_map.html

LEVEL 1

Car Park

Lecture Theatre

4

3

LT1B

LT1A

Stairs

Washroom

Executive Education Suite

Partition Walls

2 VIP / SI Room

Stairs Up to Level 2

Main Entrance

1

Sitting Lounge

Lift Lobby

---
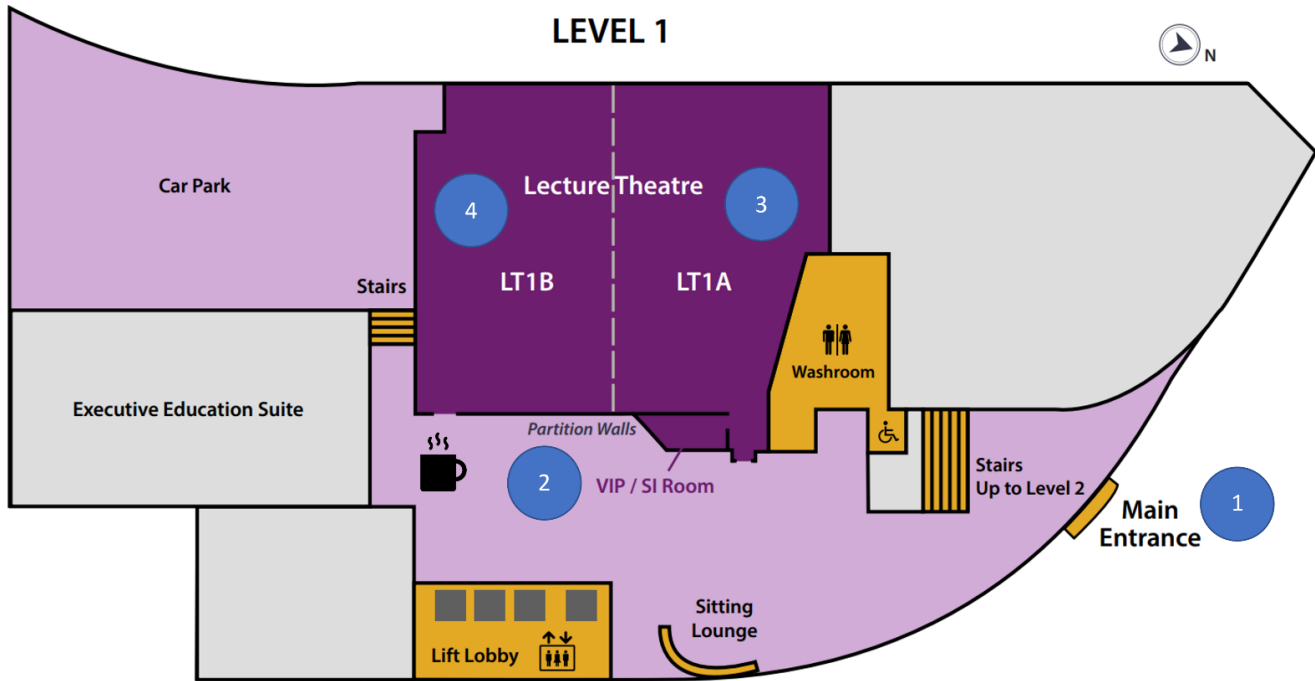
1 Entrance

2 Registration and Enquiry Counter

3 Opening Ceremony, Keynote Talks and Special Invited Session
- Lecture Theatre 1A, 1/F

4 Invited Sessions
- Lecture Theatre 1A, 1/F
- Lecture Theatre 1B, 1/F
- Room 201, 2/F
- Room 202, 2/F
- Room 203, 2/F
- Room 209A, 2/F
- Room 209B, 2/F
- Room 214, 2/F
- Room 215, 2/F
- Lecture Theatre 4, 3/F

LEVEL 2

N

Washroom

CYT 215 — 4

CYT 209B — 4

Partition Walls

CYT 214 — 4

CYT 209A — 4

CYT LT1

CYT 202 — 4

CYT 201 — 4

Washroom

CYT 213 Supporting Room

CYT 211

CYT 210

CYT 208

CYT 207

CYT 206

CYT 205

Stairs
Up to Level 3

Stairs
Down to Level 1

Sitting Lounge

CYT 203

4

Lift Lobby
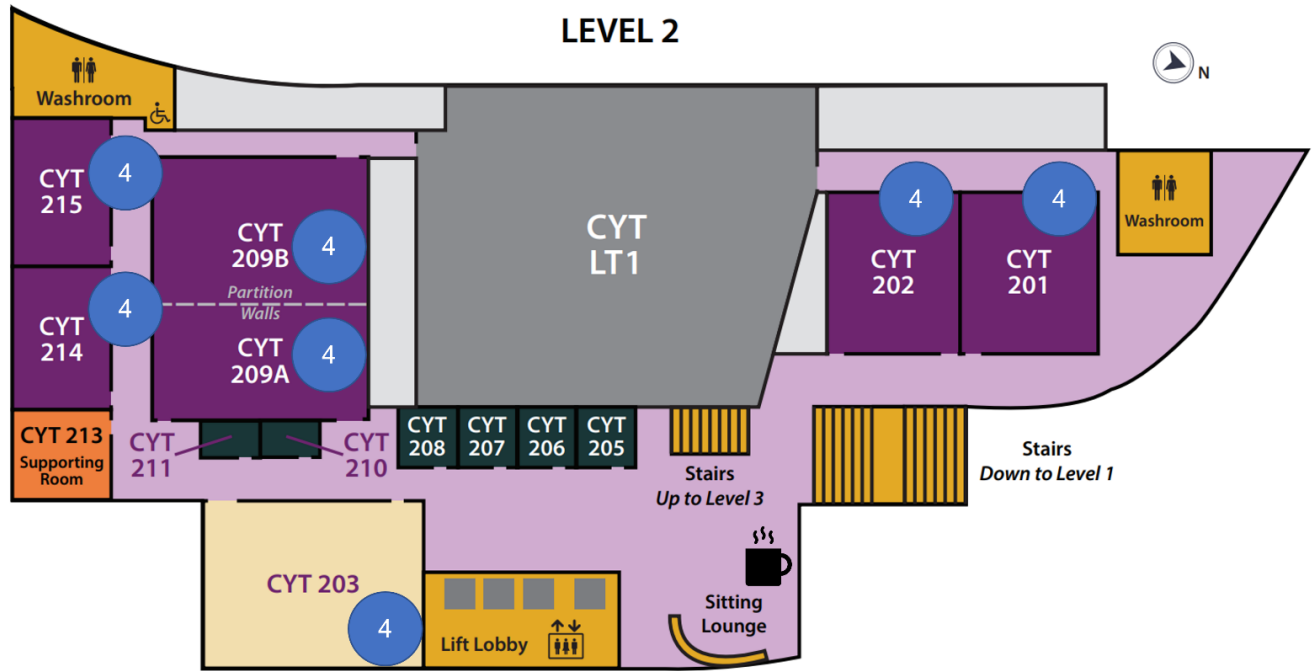
Washroom

Coffee Break

LEVEL 3
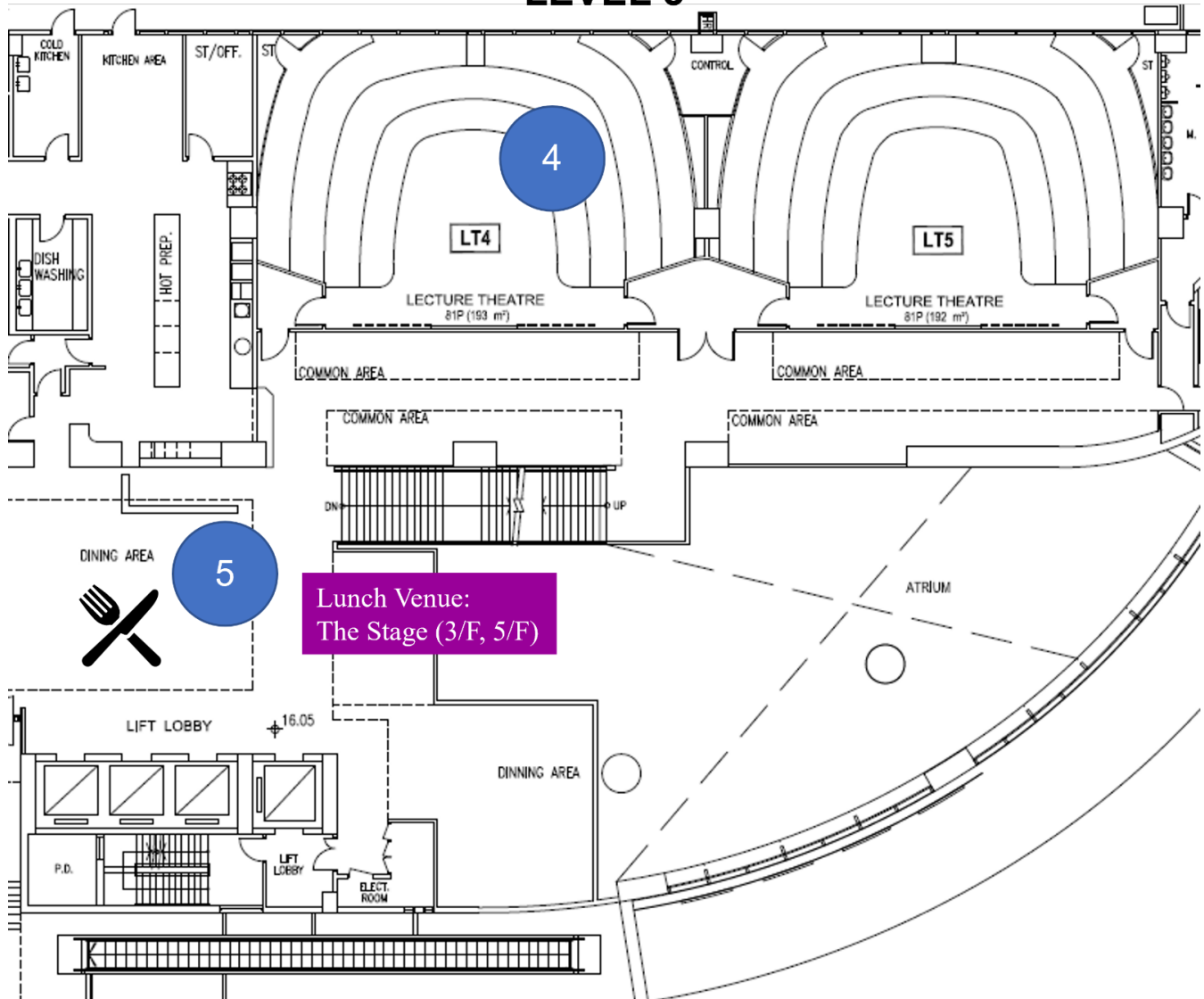
⑤ Lunch Venue 🍴
The Stage, 3/F & 5/F
* Due to limited seating in the restaurant, guests may consider having lunch before/after the lunch break (12pm-1pm) or taking the lunch boxes away and eating in their hotel rooms or in the open area. The restaurant will start serving lunch from 11:30 until 14:00. Please be reminded that drinking and eating is not allowed in the lecture theatres and classrooms.

Time (Hong Kong Time)	Activity		Host
Day 1 (July 7, 2023)			
08:30AM – 9:00AM	Registration and Enquiry from 7:30am to 6:00pm; in front of LT1A, Cheng Yu Tung Building	Welcome and Opening Ceremony Location: LT1A	Alan K.L. Chan, Jianguo (Tony) Sun, Gang Li, Xinyuan Song
09:00AM – 10:00AM		Keynote Talk 1: Songxi Chen Location: LT1A	Jianguo (Tony) Sun
10:00AM – 10:20AM		Coffee Break	
10:20AM – 12:00PM		Parallel Sessions	Session Chairs
12:00PM – 13:00PM		Lunch Break	
13:00PM – 14:40PM		Parallel Sessions	Session Chairs
14:40PM – 14:50PM		Short Break	
14:50PM – 16:30PM		Parallel Sessions	Session Chairs
16:30PM – 16:50PM		Coffee Break	
16:50PM – 18:30PM		Parallel Sessions	Session Chairs
Day 2 (July 8, 2023)			
09:00AM – 10:00AM	Registration and Enquiry from 8:00am to 6:00pm; in front of LT1A, Cheng Yu Tung Building	Keynote Talk 2: Qiman Shao Location: LT1A	Xinyuan Song
10:00AM – 10:20AM		Coffee Break	
10:20AM – 12:00PM		Parallel Sessions	Session Chairs
12:00PM – 13:00PM		Lunch Break	
13:00AM – 14:20PM		Special Invited Session Location: LT1A	Xingqiu Zhao
14:20PM – 14:30PM		Short Break	
14:30PM – 16:10PM		Parallel Sessions	Session Chairs

16:10PM – 16:30PM		Coffee Break	
16:30PM – 18:10PM		Parallel Sessions	Session Chairs
18:20 PM	Coaches will depart from the Cheng Yu Tung Building at 18:20 PM		
19:00PM – 21:00PM		Banquet	
		Day 3 (July 9, 2023)	
09:00AM – 10:00AM	Registration and Enquiry from 8:00am to 6:00pm; in front of LT1A, Cheng Yu Tung Building	Keynote Talk 3: Ji Zhu Location: LT1A	ICSA
10:00AM – 10:20AM		Coffee Break	
10:20AM – 12:00PM		Parallel Sessions	Session Chairs
12:00PM – 13:00PM		Lunch Break	
13:00AM – 14:40PM		Parallel Sessions	Session Chairs
14:40PM – 14:50PM		Short Break	
14:50PM – 16:30PM		Parallel Sessions	Session Chairs
16:30PM – 16:50PM		Coffee Break	
16:50PM – 18:30PM		Parallel Sessions	Session Chairs

Dr. Song Xi Chen is a University Chair Professor at School of Mathematical Sciences, Guanghua School of Management and Center for Statistical Science of Peking University, Peking, China. He received his B.Sc. and M.Sc. in Mathematics from Beijing Normal University in 1983 and 1988, respectively, and his Ph.D. in Statistics from Australian National University in 1993. His primary research interests include inference for high-dimensional data, environmental modeling, empirical likelihood, econometric theory and financial econometrics. He became a member of Chinese Academy of Sciences in 2021 and was elected as Fellow of the American Statistical Association and Fellow of the Institute of Mathematical Statistics (IMS) in 2009. He is also an Elected Member of International Statistical Institute, an Elected Council Member of IMS during 2016 – 2019 and an Elected Board Member of the International Chinese Statistical Association (ICSA) during 2008 – 2013. He is currently serving as Scientific Secretary of Bernoulli Society since 2019 and was selected to give Peter Hall Lecture at the 12th ICSA International Conference.

Time and Location: 9am-10am, July 7 (Hong Kong time), LT1A

Host: Jianguo (Tony) Sun

Title: Multi-level Thresholding for Detecting Rare and Faint Signals

Abstract: The work considered in this talk was largely motivated by Professor Hall works on testing for high dimensional means. I will summarize latest research on detection of rare and faint signals in the mean and the covariance matrices, the two basic summary measures of distributions, by formulating multi-level thresholding tests (MTT). The detection boundary and the minimax properties of the MTTs are presented. The MTTs are shown to be powerful in detecting sparse and weak signals in thigh dimensional mean and covariances, leading to attractive detection boundary and attain the optimal minimax rate in the signal strength under different regimes of high dimensionality and the sparsity of the signals.

Dr. Qi-Man Shao is the founding chairman of the Department of Statistics and Data Science and Chair Professor at Southern University of Science and Technology. He received his B.S in 1983 and M.Ph in 1986 from Hangzhou University (now known as Zhejiang University). Then he got his Ph.D from the University of Science and Technology of China in 1989. He was a faculty member at Hangzhou University, National University of Singapore, University of Oregon, the Hong Kong University of Science and Technology and the Chinese University of Hong Kong respectively. Dr. Shao has made significant contributions to the limit theory in probability and statistics. In particular, he has systematically developed the self-normalized limit theory and established the self-normalized large and moderate deviation theorems. In 2015, Dr. Shao was awarded the State Natural Science Award (2nd class). He was an invited speaker at the International Congress of Mathematicians in 2010 and an elected Fellow of the Institute of Mathematical Statistics (IMS) in 2001. He served as chairman of IMS Committee on Fellows and an Associate Editor of the Annals of Statistics and the Annals of Applied Probability. He is currently a council member of IMS and a co-Editor of the Annals of Applied Probability.

Time and Location: 9am-10am, July 8 (Hong Kong time), LT1A

Host: Xinyuan Song

Title: Perspective of Self-normalized Limit Theory

Abstract: Limit theory plays an important role in probability and statistics. Classical limit theorems such as the law of large numbers, the central limit theorem and the Cramér moderate deviation theorem, under deterministic standardization, have been well developed and understood. However, standardized coefficients in applications are more often random, or self-normalized. In this talk, we shall review recent developments of limit theory for self-normalized processes as well as applications to statistical inference.

Dr. Ji Zhu is Susan A. Murphy Professor of Statistics at the University of Michigan, Ann Arbor. He received his B.Sc. in Physics from Peking University, China in 1996 and M.Sc. and Ph.D. in Statistics from Stanford University in 2000 and 2003, respectively. His primary research interests include statistical machine learning, high-dimensional data modeling, statistical network analysis, and their applications to health sciences. He received an NSF CAREER Award in 2008, and was elected as a Fellow of the American Statistical Association in 2013 and a Fellow of the Institute of Mathematical Statistics in 2015. He has been rated as an ISI Highly Cited Researcher from 2014-2020 by Web of Science, which publishes an annual list recognizing leading researchers in the sciences and social sciences from around the world. He also received the International Chinese Statistical Association Pao-Lu Hsu Award in 2022.

Time and Location: 9am-10am, July 9 (Hong Kong time), LT1A

Host: Zhezhen Jin

Title: Statistical Inference on Latent Space Models for Network Data

Abstract: Recent advances in computing and measurement technologies have led to an explosion in the amount of data with network structures in a variety of fields including social networks, biological networks, transportation networks, the World Wide Web, and so on. This creates a compelling need to understand the generative mechanism of these networks and to explore various characteristics of the network structures in a principled way. Latent space models are powerful statistical tools for modeling and understanding network data. While the importance of accounting for uncertainty in network analysis is well recognized, current literature predominantly focuses on point estimation and prediction, leaving the statistical inference of latent space network models an open question. In this talk, I will present some of our recent work that aims to fill this gap by providing a general framework for analyzing the theoretical properties of the maximum likelihood estimators for latent space network models. In particular, we establish uniform consistency and individual asymptotic distribution results for latent space network models with a broad range of link functions and edge types. Furthermore, the proposed framework enables us to generalize our results to the sparse and dependent-edge scenarios. Our theories are supported by simulation studies and have the potential to be applied in downstream inferences, such as link prediction and network-assisted supervised learning.

Dr. Gang Li is a Professor of Biostatistics and Biomathematics at UCLA, is the Director of the Biostatistics and Bioinformatics Core for the UCLA Brain SPORE. He is also the Director of UCLA's Jonsson Comprehensive Cancer Center Biostatistics Shared Resource (BASE Unit).

Dr. Li is Elected Fellow of the Institute of Mathematical Statistics (2008), Elected Fellow of the American Statistical Association (2010), Elected Member of the International Statistics Institute (2000), and Elected Fellow of the Royal Statistical Society (1999). Dr. Li is Associate Editor for multiple statistics journals. His statistical research covers a broad range of areas including survival analysis, longitudinal modeling, high dimensional and large-scale data, clinical trials, and evaluation and development of biomarkers. His research papers have appeared in prestigious statistics/biostatistics journals including Annals of Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society-B, Biometrika, and Biometrics.

Besides to his methodological research, Dr. Li has collaborated extensively with basic science, translational research and clinical trial investigators. He has served as Director of Biostatistics for multiple NCI-funded P01 grants, and is a long-time collaborator with the UCLA Brain SPORE investigators. Dr. Li has been serving on UCLA's Internal Scientific Review Committee that oversee and review all cancer clinical trials at UCLA since 2007.

Time and Location: 1:00pm-1:40pm, July 8 (Hong Kong time), LT1A

Host: Xingqiu Zhao

Title: TBA

Abstract: TBA

Dr. Mei-Ling Ting Lee is a professor in the Department of Epidemiology and Biostatistics and Director of the Biostatistics and Risk Assessment Center (BRAC). Dr. Lee holds Fellowship status in several international statistical organizations, including the American Statistical Association, the Institute of Mathematical Statistics, and the Royal Statistical Society. She was named the Mosteller Statistician of the Year in 2005 by the American Statistical Association, Boston Chapter. Dr. Lee has published a book on "Analysis of Microarray Gene Expression Data" and co-edited two other books. Dr. Lee is the founding editor and editor-in-chief of the international journal Lifetime Data Analysis, the only international statistical journal that is specialized in modeling time-to-event data. The journal is currently publishing the nineteenth's volume. Dr. Lee has also co-edited three other books: Lifetime Data Models in Reliability and Survival Analysis (1995); Measurement and Statistical Analysis for Quality of Lifetime Data (2002); Risk Assessment and Evaluation of Prediction (2013).

Time and Location: 1:40pm-2:20pm, July 8 (Hong Kong time), LT1A

Host: Xingqiu Zhao

Title: Semiparametric Predictive Inference for Failure Data Using First-hitting-time Regression

Abstract: Degradation of an engineering system or disease progression in a patient can be described mathematically as a stochastic process. The system or the patient experiences a failure event when the wear and tear on the system or the patient's disease progression first reaches a critical threshold level. This happening defines a failure event and a first hitting time (FHT). First hitting time threshold regression (TR) models are based on an underlying stochastic process and hence the TR model represents a realistic alternative to the Cox model for capturing granular structure in a prediction model. To date, most applications of threshold regression have been based on parametric families of stochastic processes. This paper presents a semiparametric form of threshold regression that requires the stochastic process to have only one key property, namely, stationary independent increments. Computational aspects of the methods are straightforward. We applied the methods to data from The Osteoarthritis Initiative (OAI) study are presented to demonstrate its practical use.

Banquet Information

ClubONE Riviera (1/F, Hall A)
Date: July 8, 2023 (Saturday)
Time: 7:00pm
Address: 55-57 Tai Chung Kiu Road, Sha Tin (near Shek Mun MTR Station)
Remarks: Please bring the name badge and banquet ticket. Coaches will depart from the entrance of Cheng Yu Tung Building at 6:20pm.

水中天. (2023, March 6). In *Wikipedia*. https://zh.wikipedia.org/wiki/%E6%B0%B4%E4%B8%AD%E5%A4%A9

Sightseeing and Attractions

Hong Kong has many fascinating attractions such as Hong Kong Palace Museum, Stanley, Victoria Harbour, the Peak, Ocean Park Theme Park, Lamma Island, Disneyland, Tai O the fish village and Ngong Ping 360 Cable Car.
Please visit the website of Hong Kong Tourism Board for more details.

https://www.discoverhongkong.com/uk/index.html

ICSA 2023 International Conference Junior Researcher Award

Mingyue Du, *Postdoctoral fellow, The Hong Kong Polytechnic University*
PhD: Jilin University, June 2020

Paper: *Simultaneous Variable Selection and Estimation for Interval-Censored Failure Time Data with Ancillary Information*

Muhong Gao, *Postdoctoral Researcher, Chinese Academy of Science*
PhD: University of Wisconsin-Madison, December 2021

Paper: *Learning network-structured dependence from non-stationary multivariate point process data*

Chuoxin Ma, *Associate Professor, BNU-HKBU United International College*
PhD: University of Manchester, November 2018

Paper: *Multi-state model and structural selection for the analysis of depressive symptom dynamics in middle-aged and older adults*

Dayu Sun, *Postdoctoral Fellow, Emory University*
PhD: University of Missouri-Columbia, July 2020

Paper: *Partial Quantile Tensor Regression*

Yixin Wang, *Assistant Professor, University of Michigan*
Ph.D: Columbia University

Paper: *Desiderata for Representation Learning: A Causal Perspective*

Scientific Program (Jul. 7 - 9)

July 7, 9:00-10:00

Session 23INTKT1: Keynote Talk 1: Songxi Chen
Room: LT1A
Organizer: ICSA Committee.
Chair: Jianguo (Tony) Sun, University of Missouri-Columbia.

9:00 Multi-Level Thresholding for Detecting Rare and Faint Signals
Songxi Chen.   Peking University

9:50 Floor Discussion.

July 7, 10:20-12:00

Session 23INT90: Recent Development of Design and Analysis in Oncology Clinical Trials
Room: LT1A
Organizer: Chen Hu, Johns Hopkins University.
Chair: Ruijian Han, The Hong Kong Polytechnic University.

10:20 How Should We Compare Survival Outcomes with Non-proportional Hazards?
♦*Rick Chappell and Mitchell Paukner.*   University of Wisconsin Madison

10:45 Immune-Oncology Agents: Endpoints and Designs
Hao Wang.   Johns Hopkins University School of Medicine

11:10 On Statistical Inference of Multiple Competing Risks in Comparative Clinical Trials
Jiyang Wen, Mei-Cheng Wang and ♦*Chen Hu.*   Johns Hopkins University

11:35 Floor Discussion.

Session 23INT30: Statistical and Deep Learning for Survival Data
Room: LT1B
Organizer: Jane-Ling Wang, UC Davis.
Chair: Dayu Sun, Indiana University.

10:20 Deep Generative Estimation of Conditional Survival Function
Xingyu Zhou[1], Wen Su[2], Changyu Liu[3], Yuling Jiao[4], Xingqiu Zhao[5] and ♦*Jian Huang[5].* [1]Dow Inc. [2]The University of Hong Kong [3]The Chinese University of Hong Kong [4]Wuhan University [5]The Hong Kong Polytechnic University

10:45 Deep Learning for the Partially Linear Cox Model
♦*Qixian Zhong[1], Jonas Mueller[2] and Jane-Ling Wang[3].* [1]Xiamen University [2]Cleanlab [3]UC Davis

11:10 Statistical Inference for Counting Processes under Shape Heterogeneity
♦*Yifei Sun[1] and Ying Sheng[2].*   [1]Columbia University [2]Chinese Academy of Sciences

11:35 Variable Selection for Interval-Censored Failure Time Data
Jianguo Sun.

12:00 Floor Discussion.

Session 23INT2: Statistical Representative Points and Their Application in Statistical Inference
Room: 201
Organizer: Jiajuan Liang, BNU-HKBU United International College.
Chair: Chuoxin Ma, Beijing Normal University - HongKong Baptist University United International Collage.

10:20 Remarks on Parameter Estimation of Mixture of Two Normal Distributions Based on Small Sample Sizes
♦*Yinan Li and Kai-Tai Fang.*   BNU-HKBU United International College

10:45 Statistical Representative Points and their Application in Statistical Inference
♦*Sirao Wang, Jiajuan Liang, Min Zhou and Huajun Ye.*

11:10 On the Non-Inferiority Mcnemar Test
Zhigang Zhang.   Memorial Sloan-Kettering Cancer Center

11:35 Floor Discussion.

Session 23INT86: Statistical Methods for Robust Inference
Room: 202
Organizer: Hongyuan Cao, Florida State University.
Chair: Hongyuan Cao, Florida State University.

10:20 Residual Projection for Quantile Regression in Vertically Partitioned Big Data
Ye Fan[1], Jr-Shin Li[2] and ♦*Nan Lin[2].* [1]Capital University of Economics and Business [2]Washington University in St. Louis

10:45 Tensor Response Quantile Regression with Neuroimaging Data
Bo Wei, ♦*Limin Peng, Ying Guo, Amita Manatunga and Jennifer Stevens.*   Emory University

11:10 Recursive Quantile Estimation: Non-Asymptotic Confidence Bounds
♦*Likai Chen[1], Georg Keilbar[2] and Wei Biao Wu[3].* [1]Washington University in Saint Louis [2]University of Vienna [3]University of Chicago

11:35 An Empirical Bayes Method for Replicability Analysis of High-Dimensional Genomic Data

♦*Yan Li[1], Xiang Zhou[2], Rui Chen[3], Xianyang Zhang[4] and Hongyuan Cao[5].* [1]Jilin University [2]University of Michigan [3]Baylor College of Medicine [4]Texas A&M University [5]Florida State University

12:00 Floor Discussion.

Session 23INT76: Modern Approaches to Tackle Challenging Design Problems in Clinical Trials

Room: 203
Organizer: Weng Kee Wong, UCLA.
Chair: Weng Kee Wong, UCLA.

10:20 Seamless Phase Ii/Iii Clinical Trials with Covariate Adaptive Randomization

Hongjian Zhu.

10:45 Optimizing the Design of Pediatric Pharmacokinetic Trials – a Case Study Evaluating a Novel Drug for Treatment of Multidrug-Resistant Tuberculosis (Mdr-Tb) in Children with and without Hiv

♦*Grace Montepiedra[1], Elin Svensson[2], Weng Kee Wong[3] and Andrew Hooker[2].* [1]Harvard T.H. Chan School of Public Health [2]Uppsala University [3]UCLA

11:10 Metaheuristics for Designing Efficient Biomedical Studies

Weng Kee Wong. University of California at Los Angeles

11:35 Enhancing the Performance of Metaheuristic Algorithms by Appropriate Noise Addition

♦*Kwok Pui Choi[1], Enzio Hai Hong Kam[1], Xin Tong[1] and Weng Kee Wong[2].* [1]National University of Singapoe [2]UCLA

12:00 Floor Discussion.

Session 23INT57: Recent Developments in Optimal Treatments and High Dimensional Data Analysis

Room: 209A
Organizer: Hongyu Zhao, Yale University.
Chair: Weichuan Yu, The Hong Kong University of Science and Technology.

10:20 A Quasi-Optimal Dose-Finding Approach in Infinite Horizon Dynamic Treatment Regime

Yuhan Li[1], Wenzhuo Zhou[2] and ♦Ruoqing Zhu[1]. [1]University of Illinois Urbana Champaign [2]University of California Irvine

10:45 Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

♦*Yunan Wu[1], Lan Wang[2] and Haoda Fu[3].* [1]University of Texas at Dallas [2]University of Miami [3]Eli Lilly and Company

11:10 A Unified Generalization of Inverse Regression via Adaptive Column Selection

Yin Jin and ♦Wei Luo. Zhejiang University

11:35 Regional Quantile Regression for Multiple Responses

Eun Ryung Lee. SungKyunKwan University

12:00 Floor Discussion.

Session 23INT69: Recent Developments in Experimental Designs

Room: 209B
Organizer: Po Yang, University of Manitoba, Winnipeg, Canada.
Chair: Po Yang, University of Manitoba.

10:20 Optimal Designs in Mixed Models for Repeated Measurements

♦*Xiaojian Xu[1] and Sanjoy Sinha[2].* [1]Brock University [2]Carleton University

10:45 A Machine Learning Perspective for Optimal Design using Tight Mutual Information

♦*Xinwei Deng and Qing Guo.* Department of Statistics, Virginia Tech

11:10 A Bayesian Approach to Process Optimization on Data with Multi-Stratum Structure

Xiaohua Liu[1], ♦Po Yang[1] and Chang-Yun Lin[2]. [1]University of Manitoba, Canada [2]National Chung Hsing University, Taiwan

11:35 Generalized Bayesian d-Optimal Supersaturated Multistratum Designs

Chang-Yun Lin. NCHU

12:00 Floor Discussion.

Session 23INT7: Methodology Advances in Analyzing High-Throughput Genomic and Genetics Data

Room: 214
Organizer: Ziyi Li, The University of Texas MD Anderson Cancer Center.
Chair: Tianwei Yu, Chinese University of Hong Kong, Shenzhen (CUHK-SZ).

10:20 Summit-Fa: a New Resource for Improved Transcriptome Imputation using Functional Annotations

Melton Melton[1], Zichen Zhang[1] and ♦Chong Wu[2]. [1]Florida State University [2]UT MD Anderson Cancer Center

10:45 Individual-Specific Reference Panel Recovery Improves Cell-Type-Specific Inference

♦*Hao Feng[1], Guanqun Meng[1] and Qian Li[2].* [1]Case Western Reserve University [2]St. Jude Children's Research Hospital

11:10 A Cofunctional Grouping-Based Approach for Non-Redundant Feature Gene Selection in Unannotated Single-Cell Rna-Seq Analysis

Xiaobo Sun. Zhongnan University of Laws and Economics

11:35 Floor Discussion.

Session 23INT85: New Regression, Prediction, and Screening in Clinical Trials

Room: 215
Organizer: Catherine Liu, The Hong Kong Polytechnic University.
Chair: Catherine Liu, The Hong Kong Polytechnic University.

10:20 Covidscreen: a Web App and r Package for Assessing Asymptomatic Covid 19 Testing Strategies
♦*Li Tang, Jesse Smith, Yiwang Zhou, Motomi Mori and Akshay Sharma.* St. Jude Children's Research Hospital

10:45 Regression Analysis for Covariate-Adaptive Randomization: a Robust and Efficient Inference Perspective
♦*Wei Ma[1], Fuyi Tu[1] and Hanzhong Liu[2].* [1]Renmin University of China [2]Tsinghua University

11:10 Genomics, Machine Learning, and Fracture Risk Assessment
Qing Wu.

11:35 Low-Rank Latent Matrix-Factor Prediction Modeling for Generalized High-Dimensional Matrix-Variate Regression
Catherine Liu. The Hong Kong Polytechnic University

12:00 Floor Discussion.

Session 23INT10: Advances in Survival Analysis and Its Applications
Room: LT4
Organizer: Leilei Zeng, University of Waterloo.
Chair: Leilei Zeng, University of Waterloo.

10:20 Optimal Cut-Point of Marker for Early Disease Detection
♦*Cuiling Wang, Mindy Katz, Carol Derby and Richard Lipton.* Albert Einstein College of Medicine

10:45 Bayesian and Influence Function Based Empirical Likelihoods for Inference of Sensitivity to the Early Diseased Stage in Diagnostic Tests.
♦*Gengsheng Qin, Shuangfei Shi and Yan Hai.* Georgia State University

11:10 using Auxiliary Information for Estimation with Left Truncated Data
Yidan Shi[1], ♦Leilei Zeng[2], Mary E. Thompson[2] and Suzanne Tyas[2]. [1]University of Pennsylvania [2]University of Waterloo

11:35 Mean Residual Life Cure Models for Right-Censored Data Subject to Length-Biased Sampling
Chyong-Mei Chen[1], Hsin-Jen Chen[1] and ♦Yingwei Peng[2]. [1]National Yang Ming Chiao Tung University, Taiwan [2]Queen's University, Canada

12:00 Floor Discussion.

July 7, 13:00-14:40

Session 23INT26: Statistical Modeling of Complex-Data with Censoring and Measurement Errors
Room: LT1A
Organizer: Yanqing Sun, University of North Carolina at Charlotte, USA.
Chair: Yinghao Pan, University of North Carolina at Charlotte, USA.

13:00 Novel Empirical Likelihood Inference for the Mean Difference with Right-Censored Data
Kangni Alemdjrodo[1] and ♦Yichuan Zhao[2]. [1]Purdue University [2]Georgia State University

13:25 Semiparametric Regression Analysis of Partly Interval-Censored Failure Time Data with Application to an Aids Clinical Trial
♦*Qingning Zhou[1], Yanqing Sun[1] and Peter Gilbert[2].* [1]University of North Carolina at Charlotte [2]Fred Hutchinson Cancer Center

13:50 Analysis of Noisy Survival Data with Graphical Proportional Hazards Measurement Error Model
♦*Grace Yi[1] and Li-Pang Chen[2].* [1]University of Western Ontario [2]National Chengchi University

14:15 Floor Discussion.

Session 23INT14: Advances in Statistical Methods for Machine Learning
Room: LT1B
Organizer: Chengchun Shi, London School of Economics and Political Science.
Chair: Fan Zhou, Shanghai University of Finance and Economics.

13:00 Mdp2 Forest: a Constrained Continuous Multi-Dimensional Policy Optimization Approach for Short-Video Recommendation
Fan Zhou.

13:25 Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling
Qi Xu[1], Yubai Yuan[2], Junhui Wang[3] and ♦Annie Qu[1]. [1]UC Irvine [2]Penn State [3]CUHK

13:50 Online Statistical Inference for Matrix Contextual Bandit
Qiyu Han, ♦Will Wei Sun and Yichen Zhang. Purdue University

14:15 Floor Discussion.

Session 23INT11: Recent Developments in Survival Analysis with High-Dimensional or Longitudinal Covariates
Room: 201
Organizer: Gang Li, UCLA.
Chair: Lang Wu, UF.

13:00 Fast Lasso-Type Safe Screening for Fine-Gray Competing Risks Model with Ultrahigh Dimensional Covariates
♦*Hong Wang[1] and Gang Li[2].* [1]Central South University [2]UCLA

13:25 Sure Joint Screening for High Dimensional Cox's Proportional Hazards Model under the Case-Cohort Design
♦*Yi Liu[1] and Gang Li[2].* [1]Ocean University of China [2]University of California at Los Angeles

13:50 Kernel Meets Sieve: Transformed Hazards Models with Sparse Longitudinal Covariates

Hongyuan Cao[1], Dayu Sun[2], Zhuowei Sun[3] and ♦Xingqiu Zhao[4]. [1]Florida State University [2]Emory University [3]Jilin University [4]Hong Kong Polytechnic University

14:15 Floor Discussion.

Session 23INT103: Invited Session on Lifetime Data Analysis
Room: 202
Organizer: Mei-Ling Ting Lee, University of Maryland at College Park.
Chair: Mei-Ling Ting Lee, University of Maryland at College Park.

13:00 Kullback-Leibler-Based Relative Risk Models for Integration of Published Survival Models with New Dataset
♦*Kevin He and Di Wang.* University of Michigan

13:25 Profile Optimum Planning for Degradation Analysis
♦*Chien-Yu Peng and Ya-Shan Cheng.* Academia Sinica

13:50 Regression Analysis of Serial Gap Time Data with Recurrent and Terminal Events via Additive Hazards Models
Yong-Chen Huang and ♦Shu-Hui Chang. National Taiwan University

14:15 Surrogate Marker Assessment of Covid-19 Vaccine Efficacy using Mediation Analyses in a Case-Cohort Design
♦*Yen-Tsung Huang, Jih-Chang Yu and Jui-Hsiang Lin.* Academia Sinica

14:40 Floor Discussion.

Session 23INT34: Statistical Inference with Nonparametric and Semiparametric Methods
Room: 203
Organizer: Xinyuan Song, The Chinese University of Hong Kong.
Chair: Jingheng Cai, Sun Yat-sen University.

13:00 Paired or Partially Paired Two-Sample Tests with Unordered Samples
Yudong Wang[1], ♦Yanlin Tang[2] and Zhisheng Ye[1]. [1]National University of Singapore [2]East China Normal University

13:25 Estimation and Inference for Ultra-High Dimensional Quasi-Likelihood Models Based on Data Splitting
Xuejun Jiang. Southern University of Science and Technology

13:50 Dimension Reduction for Functional Time Series Model
♦*Guochang Wang and Zenyao Wen.*

14:15 Regression Analysis of Partially Linear Transformed Mean Residual Life Models
♦*Haijin He[1], Jingheng Cai[2] and Xinyuan Song[3].* [1]Shenzhen University [2]Sun-Yat sen University [3]The Chinese University of Hong Kong

14:40 Floor Discussion.

Session 23INT8: High-Dimensional Data Analysis: Classification and Testing
Room: 209A
Organizer: Xuejun JIANG, Department of Statistics and Data Science, Southern University of Science and Technology.
Chair: Xuejun JIANG, Department of Statistics and Data Science, Southern University of Science and Technology.

13:00 A Powerful Methodology for Analyzing Correlated High Dimensional Data with Factor Models
Peng Wang[1], Pengfei Lyu[2], Shyamal Peddada[3] and ♦Hongyuan Cao[2]. [1]Jilin University [2]Florida State University [3]NIEHS

13:25 Multi-Threshold Structural Equation Model
Jingli Wang. Nankai University

13:50 Empirical Likelihood Ratio Tests for Nonnested Model Selection Based on Predictive Losses
Jiancheng Jiang[1], Xuejun Jiang[2] and ♦Haofeng Wang[2]. [1]University of North Carolina at Charlotte [2]Southern University of Science and Technology

14:15 Asymptotic Normality for Eigenvalue Statistics of a General Sample Covariance Matrix when $p/n \to \infty$ and Applications
Jiaxin Qiu[1], ♦Zeng Li[2] and Jianfeng Yao[3]. [1]The University of Hong Kong [2]Southern University of Science and Technology [3]Chinese University of Hong Kong, Shenzhen

14:40 Floor Discussion.

Session 23INT71: Recent Advances in Functional Data Analysis
Room: 209B
Organizer: Jie Li, School of Statistics, Renmin University of China.
Chair: Jie Li, School of Statistics, Renmin University of China.

13:00 Testing Linearity in Semi-Functional Partially Linear Regression Models
♦*Yongzhen Feng[1], Jie Li[2] and Xiaojun Song[3].* [1]Tsinghua University [2]Renmin University of China [3]Peking University

13:25 Time-Varying Treatment Effects of Functional Data with Latent Confounders: Application to Sleep Heart Health Studies
♦*Jie Li[1], Shujie Ma[2] and Yehua Li[2].* [1]Renmin University of China [2]University of California, Riverside

13:50 Ftir Feature Extraction with Forensic Microtraces Combining Screening and Functional Discriminant Analysis
Xing Wang. School of Statistics, Renmin University of China

14:15 Network Vector Autoregression with Time-Varying Nodal Influence
♦*Yi Ding[1], Rui Pan[2] and Bo Zhang.* [1]Renmin University of China [2]Central University of Finance and Economics

14:40 Floor Discussion.

Session 23INT13: Recent Developments on Causal Inference and Genetics

Room: 214
Organizer: Zhonghua Liu, Columbia University.
Chair: Zhonghua Liu, Columbia University.

13:00 Staarpipeline: An all-in-One Rare-Variant Tool for Biobank-Scale Whole-Genome Sequencing Data
♦*Zilin Li*[1], *Xihao Li*[2] *and Xihong Lin*[2]. [1]Indiana University School of Medicine [2]Harvard T.H. Chan School of Public Health

13:25 Causal Inference with Invalid Instruments: Post-Selection Problems and a Solution using Searching and Sampling
Zijian Guo. Rutgers

13:50 Optimal Individualized Decision-Making with Proxies
Tao Shen[1] *and* ♦*Yifan Cui*[2]. [1]NUS [2]ZJU

14:15 Recent Progress in Machine Learning-Based Causal Inference
Lin Liu. Shanghai Jiao Tong University

14:40 Floor Discussion.

Session 23INT31: New Development on Precision Medicine

Room: 215
Organizer: Ling Zhou, Southwestern University of Finance and Economics.
Chair: Ling Zhou, Southwestern University of Finance and Economics.

13:00 Dynamic Logistic State Space Prediction Model for Clinical Decision Making
♦*Jiakun Jiang*[1], *Wei Yang*[2], *Erin M. Schnellinger*[2], *Stephen E. Kimmel*[2] *and Wensheng Guo*[2]. [1]Beijing Normal University [2]University of Pennsylvania

13:25 Center-Augmented L_2-Type Regularization for Subgroup Learning
♦*Ye He*[1], *Ling Zhou*[2], *Yingcun Xia*[3] *and Huazhen Lin*[2]. [1]Sichuan Normal University,China [2]Southwestern University of Finance and Economics,China [3]National University of Singapore, Singapore

13:50 Generalized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types
♦*Wei Liu*[1], *Huazhen Lin*[2], *Shurong Zheng*[3] *and Jin Liu*[4]. [1]Centre for Quantitative Medicine, Program in Health Services & Systems Research, Duke-NUS Medical School [2]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics [3]School of Mathematics and Statistics, Northeast Normal University [4]School of Data Science, The Chinese University of Hong Kong-Shenzhen

14:15 Subgroup-Effects Models for the Analysis of Personal Treatment Effects
♦*Ling Zhou*[1], *Shiquan Sun*[2], *Haoda Fu*[3] *and Peter Song*[4]. [1]Southwestern University of Finance and Economics [2]Xi'an Jiaotong University [3]Eli Lilly and Company [4]University of Michigan

14:40 Floor Discussion.

July 7, 14:50-16:30

Session 23INT18: High-Dimensional Regression, State Space Models, and COVID-19 Prediction

Room: LT1A
Organizer: Yuedong Wang, University of California - Santa Barbara.
Chair: Yuedong Wang, University of California - Santa Barbara.

14:50 Sparse Convoluted Rank Regression in High Dimensions
♦*Le Zhou*[1], *Boxiang Wang*[2] *and Hui Zou*[3]. [1]Hong Kong Baptist University [2]University of Iowa [3]University of Minnesota

15:15 Dynamic Hierarchical State Space Forecasting
♦*Ziyue Liu*[1] *and Wensheng Guo*[2]. [1]Indiana University School of Medicine [2]University of Pennsylvania

15:40 Predicting Sars-Cov-2 Infection among Hemodialysis Patients using Multimodal Data
♦*Juntao Duan*[1], *Hanmo Li*[1], *Xiaoran Ma*[1], *Yuedong Wang*[1], *Peter Kotanko*[2], *Hanjie Zhang*[3], *Rachel Lasky*[4], *Caitlin Monaghan*[4], *Mengyang Gu*[1] *and Wensheng Guo*[5]. [1]Department of Statistics and Applied Probability, University of California, Santa Barbara, California, United States [2]Icahn School of Medicine at Mount Sinai, New York, United States [3]Renal Research Institute, New York, United States [4]Fresenius Medical Care, Global Medical Office, Waltham, Massachusetts [5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States

15:05 Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with Covid-19
Xiaoran Ma[1], *Wensheng Guo*[2], *Mengyang Gu*[1], *Peter Kotanko*[3], *Len Usvyat*[4] *and* ♦*Yuedong Wang*[1]. [1]University of California - Santa Barbara [2]University of Pennsylvania [3]Renal Research Institute [4]Fresenius Medical Care

16:30 Floor Discussion.

Session 23INT12: Recent Developments in Analysis of Functional, Longitudinal, and Time-to-Event Data

Room: LT1B
Organizer: Gang Li, UCLA.
Chair: Xingqiu Zhao, The Hong Kong Polytechnic University.

14:50 Functional Data Analysis with Covariate-Dependent Mean and Covariance Structures*
Huazhen Lin. Southwestern University of Finance and Economics

15:15 Robust Inference for Joint Models of Longitudinal and Survival Data
Lang Wu. University of British Columbia, Vancouver

15:40 Functional Data Modeling in High Dimensions: Fundamentals, Sparsity and Fast Computation.
Shaojun Guo.

15:05 Floor Discussion.

Session 23INT94: Advanced Statistical Methods for Complex Observational Studies and Clinical Trials
Room: 201
Organizer: Jianwen Cai, University of North Carolina at Chapel Hill.
Chair: Yuanshan Wu, Zhongnan University of Economics and Law.

14:50 Survival Analysis of Randomized Controlled Trials with Auxiliary Patient Population Information from Observational Studies
Xiaofei Wang. Duke University

15:15 Subgroup Analysis for Longitudinal Data Based on a Partial Linear Varying Coefficient Model with a Change Plane
Guoyou Qin. Fudan University

15:40 Practical Considerations in Trial Design with Win Ratio Method for Multiple Time-to-Event Endpoints with Hierarchy
♦*Huiman Barnhart, Yuliya Lokhnygina, Roland Matsouaka and Frank Rockhold.* Duke University

15:05 Accelerated Failure Time Modeling via Nonparametric Mixtures
Byungtae Seo[1] *and* ♦*Sangwook Kang*[2]. [1]Sungkyunkwan University [2]Yonsei University

16:30 Floor Discussion.

Session 23INT19: Advanced Statistical Learning Methods for Heterogeneous Data and Model Integration
Room: 202
Organizer: Lu Tang, University of Pittsburgh.
Chair: Peter Song, University of Michigan.

14:50 Evaluation of Combined Data from Subgroup Selection and Validation Phases in Clinical Trials
♦*Xinzhou Guo*[1], *Jianjun Zhou*[2] *and Xuming He*[3]. [1]Hong Kong University of Science and Technology [2]Yunnan University [3]University of Michigan

15:15 d-Gcca: Decomposition-Based Generalized Canonical Correlation Analysis for Multi-View High-Dimensional Data
♦*Hai Shu*[1], *Zhe Qu*[2] *and Hongtu Zhu*[3]. [1]New York University [2]Tulane University [3]The University of North Carolina at Chapel Hill

15:40 Robust Transfer Learning of Individualized Treatment Rules
Lu Tang. University of Pittsburgh

15:05 Floor Discussion.

Session 23INT67: Novel Statistical Models and Methods with Applications
Room: 203
Organizer: Hon Keung Tony Ng, Bentley University.
Chair: Man Ho Ling, The Education University of Hong Kong.

14:50 Bayesian Inference and Applications for Zero-Inflated Models
Seong Kim. Hanyang University

15:15 The Gender Gap in Venture Capital Market: a Statistical Approach using Structural Matching Models and Accelerator Data
Chuan Chen[1] *and* ♦*Junnan He*[2]. [1]Wisconsin School of Business [2]Sciences Po

15:40 Semiparametric Evaluation of First-Passage Distribution for Step-Stress Accelerated Degradation Tests
Lochana Palayangoda[1], ♦*Hon Keung Tony Ng*[2] *and Ling Li*[3]. [1]Department of Mathematical and Statistical Sciences, University of Nebraska Omaha [2]Department of Mathematical Sciences, Bentley University [3]Xi'an Microelectronic Technology Institute

15:05 Floor Discussion.

Session 23INT58: Statistical Genetics and Genomics
Room: 209A
Organizer: Hongyu Zhao, Yale University.
Chair: Eun Ryung Lee.

14:50 Transfer Learning for High-Dimensional Multiple Response Regression
Seyoung Park. Sungkyunkwan University

15:15 Decoding Gene Functions: Exploring their Significance in Biological Context
Ying Zhu. Fudan University

15:40 Multi-Tissue Transcriptome-Wide Association Studies with High Dimensional Transfer Learning
♦*Daoyuan Lai, Han Wang and Yan Dora Zhang.* Department of Statistics and Actuarial Science, The University of Hong Kong

15:05 A Statistical Framework for Cross-Population Fine-Mapping by Leveraging Genetic Diversity and Accounting for Confounding Bias
♦*Mingxuan Cai*[1], *Zhiwei Wang*[2], *Jiashun Xiao*[3], *Xianghong Hu*[2], *Gang Chen*[4] *and Can Yang*[2]. [1]City University of Hong Kong [2]The Hong Kong University of Science and Technology [3]Shenzhen Research Institute of Big Data [4]The WeGene Company

16:30 Floor Discussion.

Session 23INT88: Recent Advances in Statistical Methods and Machine Learning.
Room: 209B
Organizer: Yuanyuan Lin, The Chinese University of Hong Kong.
Chair: Wenxin Zhou, University of California, San Diego.

14:50 Super-Quantile Regression
Xuming He[1], *Kean Ming Tan*[2] *and* ♦*Wenxin Zhou*[3]. [1]Washington University in St. Louis [2]University of Michigan [3]University of Illinois at Chicago

15:15 Complexity of Feed-Forward Neural Networks from the Perspective of Functional Equivalence
Guohao Shen. The Hong Kong Polytechnic University

15:40 Floor Discussion.

Session 23INT20: Statistical Methods and Applications in Precision Medicine
Room: 214
Organizer: Xiaoqing Pan, Shanghai Normal University.
Chair: Xiaoqing Pan, Shanghai Normal University.

14:50 Wemics: a Single-Base Resolution Methylation Quantification Method by Weighting Methylation of Consecutive Cpg Sites
Yi Liu. Zhejiang University

15:15 Robust Method for Optimal Treatment Decision Making Based on Survival Data
♦*Yuexin Fang*[1], *Baqun Zhang*[2] *and Min Zhang*[3]. [1]Shanghai Normal University [2]Shanghai University of Finance and Economics [3]University of Michigan

15:40 Methods for Identifying Differentially Methylated Regions for Monozygotic Twins
♦*Xiaoqing Pan*[1], *Pengyuan Liu*[2], *Srividya Kidambi*[3] *and Mingyu Liang*[3]. [1]Shanghai Normal University [2]Zhejiang University [3]Medical College of Wisconsin

15:05 Driverwmds: a Powerful Network Control Method for Predicting Cancer Driver Genes
Xiang Cheng[1], *Xiaoqing Pan*[2], *Pengyuan Liu*[1] *and* ♦*Yan Lu*[1]. [1]Zhejiang University [2]Shanghai Normal University

16:30 Floor Discussion.

Session 23INT27: Inference for High Dimensional Data
Room: 215
Organizer: Likai Chen, Washington University in Saint Louis.
Chair: Likai Chen, Washington University in Saint Louis.

14:50 Cp Factor Model for Dynamic Tensors
♦*Yuefeng Han*[1], *Cun-Hui Zhang*[2] *and Rong Chen*[2]. [1]University of Notre Dame [2]Rutgers University

15:15 L_2 Inference of High Dimensional Change Points Detection
Weining Wang. University of York

15:40 Online Change Point Detection in High-Dimensional Factor Models
♦*Mengyu Xu and Mahdi Mirhosseini.* University of Central Florida

15:05 Floor Discussion.

Session 23INT16: Statistical Methods for Complex Medical Data
Room: LT4
Organizer: Ruoqing Zhu, University of Illinois Urbana Champaign.
Chair: Yifan Cui, Zhejing University.

14:50 Integrative Factor Regression and Its Inference for Multimodal Data Analysis
Quefeng Li. UNC, Chapel Hill

15:15 Bayesian Analysis for Imbalanced Positive-Unlabelled Diagnosis Codes in Electronic Health Records
Ru Wang[1], ♦*Ye Liang*[2], *Zhuqi Miao*[3] *and Tieming Liu*[2]. [1]Dell Inc. [2]Oklahoma State University [3]State University of New York at New Paltz

15:40 Truncation Model Analysis for the under-Reporting Probability in Covid-19 Pandemic
♦*Wei Liang, Hongsheng Dai and Marialuisa Restaino.*

15:05 Doubly Robust Methods for Selecting Optimal Treatment Based on Observational Data
Qian Xu, ♦*Qi Zheng and Maiying Kong.* University of Louisville

16:30 Floor Discussion.

July 7, 16:50-18:30

Session 23INT32: Real-World Challenges and Recent Developments of Statistics in Biosciences
Room: LT1A
Organizer: Joan Hu, Simon Fraser University, Canada.
Chair: Trevor Thomson, Simon Fraser University, Canada.

16:50 Data Integration and Subsampling Techniques in Distribution Estimation for Event Times with Missing Origins
♦*Yi Xiong*[1] *and Joan Hu*[2]. [1]University of Manitoba; Fred Hutchinson Cancer Center [2]Simon Fraser University

17:15 A Latent Variable Cace Model for Multidimensional Endpoints and Treatment Noncompliance with Application to a Longitudinal Trial of Arthritis Health Journal
♦*Lulu Guo*[1], *Joan Hu*[1], *Yi Qian*[2], *Diane Lacaille*[2] *and Hui Xie*[1]. [1]Simon Fraser University, Canada [2]University of British Columbia, Canada

17:40 A Step-Wise Multiple Testing for Linear Regression Models with Application to the Study of Resting Energy Expenditure
Junyi Zhang[1], *Zimian Wang*[2], ♦*Zhezhen Jin*[3] *and Zhiliang Ying*[4]. [1]Paul H. Chook Department of Information Systems and Statistics, Baruch College, 55 Lexington Avenue, New York, NY 10010, USA [2]Obesity Research Center, Columbia University, New York, NY 10025, USA [3]Department of Biostatistics, Columbia University, New York, NY 10032, USA [4]Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

17:05 Data Masking with Random Orthogonal Matrix
Samuel Wu.

18:30 Floor Discussion.

Session 23INT38: Some Modern Issues of Statistical Learning
Room: LT1B
Organizer: Xiaodan Fan, The Chinese University of Hong Kong.
Chair: Tiejun Tong, Hong Kong Baptist University.

16:50   From Genetic Correlation Analysis to Cross-Ancestry Genetic Prediction

Lin Hou.   Tsinghua University

17:15   An Integrated Deep Learning Framework for the Interpretation of Untargeted Metabolomics Data

Leqi Tian and ♦Tianwei Yu.

17:40   Robust Statistical Learning on Heavy Tailed Data

♦*Lihu Xu*[1]*, Fang Xiao*[2]*, Qiuran Yao*[1] *and Huiming Zhang*[1]*.* [1]University of Macau [2]Peking University

17:05   Time-Series Forecasting of Mortality Rates using Transformer

Chaojie Wang.   Jiangsu University

18:30   Floor Discussion.

Session 23INT36: New Challenges in Modelling High-Dimensional and Complex Data
Room: 201
Organizer: Xinyuan Song, The Chinese University of Hong Kong.
Chair: Yanlin Tang, East China Normal University.

16:50   Multi-Kink Quantile Regression for Longitudinal Data with Application to Progesterone Data Analysis

Chuang Wan[1]*, ♦Wei Zhong*[1]*, Wenyang Zhang*[2] *and Changliang Zou*[3]*.* [1]Xiamen University [2]The University of York [3]Nankai University

17:15   Factor Modeling for High-Dimensional Time Series with Single-Index Factor Loadings

Tao Huang.   Shanghai University of Finance and Economics

17:40   Functional Data Analysis for Earth Observation

Julian Austin[1]*, Robin Henderson*[1] *and ♦Jian Qing Shi*[2]*.* [1]Newcastle University [2]Southern University of Science and Technology

17:05   Floor Discussion.

Session 23INT47: Recent Advances in Reliability
Room: 202
Organizer: Ping Shing Ben Chan, The Chinese University of Hong Kong.
Chair: Man Ho Ling, The Education University of Hong Kong.

16:50   Assessing Cyber Risks of Networked Systems Based on l-Hop Propagation

Gaofeng Da.   Nanjing University of Aeronautics and Astronautics

17:15   Non-Parametric Inference Based on Reliability Life-Test of Non-Identical Coherent Systems

Xiaojun Zhu.   Xi'an Jiaotong-Liverpool University

17:40   Dependent Stress–strength Reliability of Multi-State System by Copulas using Improved Generalized Survival Signature

Xuchao Bai[1] *and ♦Mu He*[2]*.* [1]Xidian University [2]Xi'an Jiaotong-Liverpool University

17:05   Floor Discussion.

Session 23INT15: Advances in Health and Lifetime Data Science
Room: 203
Organizer: Shu-Hui Chang, National Taiwan University.
Chair: Shu-Hui Chang, National Taiwan University.

16:50   A Tree-Based Bayesian Accelerated Failure Time Cure Model for Estimating Heterogeneous Treatment Effect

Rongqian Sun and ♦Xinyuan Song.   The Chinese University of Hong Kong

17:15   Optimizing Svm Parameters for High-Dimensional Class-Imbalanced Data Sets

Chen-An Tsai.   National Taiwan University

17:40   Synthesizing Auxiliary Information in Analyzing Survival Data with Population Heterogeneity

♦*Yu-Jen Cheng*[1]*, Yen-Chun Liu*[2]*, Chang-Yu Tsai*[1] *and Chiung-Yu Huang*[3]*.* [1]National Tsing Hua University [2]Duke University [3]University of California at San Francisco

17:05   On a Surrogate Measure for Time-Varying Biomarkers in Randomized Clinical Trials

Ying Qing Chen.   Stanford University

18:30   Deep Convolutional Neural Networks for Multiclass Classification of Three Dimensional Brain Images

♦*Guan-Hua Huang, Chih-Hsuan Lin and Yu-Ren Cai.* National Yang Ming Chiao Tung University

Floor Discussion.

Session 23INT5: Recent Progresses on Change-Point Analysis
Room: 209A
Organizer: Hao Chen, University of California, Davis.
Chair: Mu Yue, Singapore University of Technology & Design (SUTD).

16:50   Online Kernel Cusum for Change-Point Detection

Song Wei and ♦Yao Xie.   Georgia Institute of Technology

17:15   Detecting Multiple Anomaly Regions on Spatial Grid

♦*Chao Zheng and Baiyu Wang.*   University of Southampton

17:40   Likelihood Score Method for Change-Points Estimation on Large-Scale Data Streams

♦*Shouri Hu*[1]*, Jingyan Huang*[2]*, Hao Chen*[3] *and Hock Peng Chan*[2]*.* [1]University of Electronic Science and Technology of China [2]National University of Singapore [3]University of California at Davis

17:05 Optimal Difference-Based Variance Estimation in Change Point Analysis and Trend Inference
Kin Wai Chan. The Chinese University of Hong Kong

18:30 Floor Discussion.

Session 23INT65: Statistical Methods and Applications in High Dimensional Biological Data
Room: 209B
Organizer: Xuekui Zhang, University of Victoria.
Chair: Xuekui Zhang, University of Victoria.

16:50 Multi-Task Prediction Model for Time to Event Data
Shuai You[1], Xiaowen Cao[1], Grace Yi[2], ♦Xuekui Zhang[1] and Li Xing[3]. [1]University of Victoria [2]University of Western Ontario [3]University of Saskatewan

17:15 Clustering Single-Cell Rna Sequencing Data using a Mixture Model-Based Deep Learning Algorithm
Leann Lac[1], Eric Lin[2], Boyuan Liu[2], Daryl Fung[1], Carson Leung[1] and ♦Pingzhao Hu[3]. [1]University of Manitoba [2]University of Toronto [3]Western University

17:40 Group Effects in Linear Models
Min Tsao. University of Victoria, Canada

17:05 Floor Discussion.

Session 23INT17: New Statistical Methods for Analyzing Complex Survival Data
Room: 214
Organizer: Qingning Zhou, Assistant Professor, University of North Carolina at Charlotte.
Chair: Qingning Zhou, Assistant Professor, University of North Carolina at Charlotte.

16:50 Longitudinal Variable Selection under Joint Modeling of Longitudinal and Survival Data
♦Fangfang Wang[1], Lu Lin[2], Lei Liu[3], Hongmei Jiang[4] and Lihui Zhao[4]. [1]Yancheng Institute of Technology [2]Shandong University [3]Washington University in St. Louis [4]Northwestern University

17:15 Scalable Estimation for High Velocity Survival Data
♦Ying Sheng[1], Yifei Sun[2], Charles E. Mcculloch[3] and Chiung-Yu Huang[3]. [1]Chinese Academy of Sciences [2]Columbia University [3]University of California at San Francisco

17:40 A Semiparametric Cox-Aalen Transformation Model with Censored Data
Xi Ning[1], Yinghao Pan[2], ♦Yanqing Sun[2] and Peter Gilbert[3]. [1]University of North Carolina at Charlotte [2]University of North Carolina at Charlotte, USA [3]Fred Hutchinson Cancer Center and University of Washington

17:05 Conditional Quasi-Likelihood Inference for Mean Residual Life Regression with Clustered Failure Time Data
Rui Huang and ♦Liming Xiang. Nanyang Technological University, Singapore

18:30 Floor Discussion.

Session 23INT46: Recent Developments in Statistical Machine Learning
Room: 215
Organizer: Will Wei Sun, Purdue University.
Chair: Will Wei Sun, Purdue University.

16:50 Regularized Greedy Gradient q-Learning with Mobile Health Applications.
Min Qian. Columbia University

17:15 Differential Privacy in Personalized Pricing with Nonparametric Demand Models
Xi Chen[1], Sentao Miao[2] and ♦Yining Wang[3]. [1]New York University [2]McGill University [3]University of Texas at Dallas

17:40 Learning Linear Non-Gaussian Dag with Diverging Number of Nodes
Junhui Wang. Chinese University of Hong Kong

17:05 Online Estimation with Dependent Samples and Robust Policy Evaluation in Reinforcement Learning
Xi Chen[1], Weidong Liu[2], Jiyuan Tu[2] and ♦Yichen Zhang[3]. [1]New York University [2]Shanghai Jiao Tong University [3]Purdue University

18:30 Floor Discussion.

July 8, 9:00-10:00

Session 23INTKT2: Keynote Talk 2: Qiman Shao
Room: LT1A
Organizer: ICSA Committee.
Chair: Xinyuan Song, The Chinese University of Hong Kong.

9:00 Perspective of Self-Normalized Limit Theory
Qi-Man Shao. Southern University of Science and Technology

9:50 Floor Discussion.

July 8, 10:20-12:00

Session 23INT100: Network Modeling and Applications
Room: LT1A
Organizer: Jianqing Fan, Princeton University.
Chair: Jianqing Fan, Princeton University.

10:20 Statistics About Statisticians
Jiashun Jin.

10:45 Individual-Centered Partial Information in Social Networks
Xiao Han[1], Rachel Wang[2] and ♦Xin Tong[3]. [1]University of Science and Technology of China [2]University of Sydney [3]University of Southern California

11:10 Group Network Hawkes Process
Guanhua Fang[1], Ganggang Xu[2], Haochen Xu[1], ♦Xuening Zhu[1] and Yongtao Guan[2]. [1]Fudan University [2]University of Miami

11:35 Inference for Heteroscedastic Pca with Missing Data
♦*Jianqing Fan and Yihong Gu.*   Princeton University

12:00 Floor Discussion.

Session 23INT49:   Recent Developments in Deep Learning
Room: LT1B
Organizer: Jian Huang, The Hong Kong Polytechnic University.
Chair: Jian Huang, The Hong Kong Polytechnic University.

10:20 Conditional Stochastic Interpolation: a New Approach to Conditional Sampling
♦*Ding Huang, Guohao Shen, Ting Li and Jian Huang.*
The Hong Kong Polytechnic University

10:45 Robust Structure Learning and L_p-Regularization for Graph Neural Networks.
Shaogao Lv.

11:10 Optimal Rates of Approximation by Shallow ReluǨ Neural Networks and Applications to Nonparametric Regression
♦*Yunfei Yang*[1] *and Ding-Xuan Zhou*[2].   [1]City University of Hong Kong [2]University of Sydney

11:35 Deep Sufficient Representation Learning via Mutual Information
♦*Siming Zheng*[1]*, Yuanyuan Lin*[1] *and Jian Huang*[2].   [1]The Chinese University of Hong Kong [2]The Hong Kong Polytechnic University

12:00 Error Analysis for Deep Adversarial Traning
Yuling Jiao.

Floor Discussion.

Session 23INT44:   Recent Advances in Matrix and Tensor Data Analysis
Room: 201
Organizer: Emma Jingfei Zhang, Emory Univeristy.
Chair: Yingying Fan, University of Southern California.

10:20 Inference for Heteroskedastic Pca with Missing Data
Yuling Yan[1]*,* ♦*Yuxin Chen*[2] *and Jianqing Fan*[1].
[1]Princeton University [2]University of Pennsylvania

10:45 Tensor t Distribution and Tensor Response Regression
♦*Ning Wang*[1]*, Qing Mai*[2] *and Xin Zhang*[2].   [1]Beijing Normal University [2]Florida State University

11:10 Matrix Completion with Model-Free Weighting
Jiayi Wang[1]*,* ♦*Raymond K. W. Wong*[1]*, Xiaojun Mao*[2] *and Kwun Chuen Gary Chan*[3].   [1]Texas A&M University [2]Fudan University [3]University of Washington

11:35 Floor Discussion.

Session 23INT96:   Recent Developments in Genetics and Genomics and High Dimensional Data
Room: 202
Organizer: Laura Zhou, University of North Carolina at Chapel Hill.
Chair: Xiaofei Wang, Duke University.

10:20 Improving Polygenic Risk Prediction in Admixed Populations by Explicitly Modeling Ancestral-Specific Effects via Gaudi
Yun Li.   University of North Carolina at Chapel Hill

10:45 Machine-Learning-Based Genotype Imputation Quality Calibration
♦*Quan Sun*[1]*, Yingxi Yang*[2]*, Jonathan Rosen*[1]*, Laura Raffield*[1]*, Michael Bamshad*[3]*, Garry Cutting*[4]*, Michael Knowles*[1]*, Daniel Schrider*[1]*, Christian Fuchsberger*[5] *and Yun Li*[1].   [1]University of North Carolina at Chapel Hill [2]Yale University [3]University of Washington [4]Johns Hopkins University [5]EURAC Research

11:10 High-Dimensional Robust Inference for Censored Linear Models
Jiayu Huang[1] *and* ♦*Yuanshan Wu*[2].   [1]Wuhan University [2]Zhongnan University of Economics and Law

11:35 Statistical Method for Tct Data
Si Liu, Phil Bradley and ♦*Wei Sun.*   Fred Hutchinson Cancer Center

12:00 Floor Discussion.

Session 23INT28: Functional and Metric Space Data
Room: 203
Organizer: Jane-Ling Wang, UC Davis.
Chair: Jun Zhao, Rutgers University.

10:20 Dimension Reduction for Frechet Regression
Qi Zhang, Lingzhou Xue and ♦*Bing Li.*   Pennsylvania State University

10:45 A Unified Approach to Hypothesis Testing for Functional Linear Models
Yinan Lin and ♦*Zhenhua Lin.*   National University of Singapore

11:10 Geometric Eda for Random Objects
♦*Paromita Dubey*[1]*, Yaqing Chen*[2] *and Hans-Georg Müller*[3].   [1]USC [2]Rutgers University [3]UC Davis

11:35 Geometric Exploration of Random Objects Through Optimal Transport
Paromita Dubey[1]*,* ♦*Yaqing Chen*[2] *and Hans-Georg Müller*[3].   [1]University of Southern California [2]Rutgers University [3]University of California, Davis

12:00 Floor Discussion.

Session 23INT48: Recent Developments in Statistical Genomics with Applications to COVID-19
Room: 209A
Organizer: Yingying Wei, The Chinese University of Hong Kong.
Chair: Yingying Wei, The Chinese University of Hong Kong.

10:20 Differential Inference for Single-Cell Rna-Sequencing Data
♦*Fangda Song*[1]*, Kevin Yip*[2] *and Yingying Wei*[2].   [1]The Chinese University of Hong Kong, Shenzhen [2]The Chinese University of Hong Kong

10:45 Identification of Cell-Type-Specific Spatially Variable Genes Accounting for Excess Zeros

Xiangyu Luo. Renmin University of China

11:10 A Novel Penalized Inverse-Variance Weighted Estimator for Mendelian Randomization with Applications to Covid-19 Outcomes

Siqi Xu[1], Peng Wang[2], Wing Kam Fung[1] and ♦Zhonghua Liu[3]. [1]University of Hong Kong [2]Huazhong University of Science and Technology [3]Columbia University

11:35 Subtyping of Major Sars-Cov-2 Variants Reveals Different Transmission Dynamics Based on 10 Million Genomes

♦*Hsin-Chou Yang[1], Jen-Hung Wang[1], Chih-Ting Yang[1], Yin-Chun Lin[1], Han-Ni Hsieh[1], Po-Wen Chen[1], Hsiao-Chi Liao[1], Chun-Houh Chen[1] and James C. Liao[2].* [1]Institute of Statistical Science, Academia Sinica [2]Institute of Biological Chemistry, Academia Sinica

12:00 Floor Discussion.

Session 23INT61: Recent Developments on Statistical Inference and Clustering

Room: 209B
Organizer: Xiaotong Shen, University of Minnesota.
Chair: Junhui Wang, City University of Hong Kong.

10:20 Inferring Social Influence in Dynamic Networks

Yuguo Chen. University of Illinois Urbana-Champaign

10:45 Bayesian Biclustering and Its Application in Education Data Analysis

Weining Shen.

11:10 Bootstrap the Cross-Validation Estimator

Bryan Cai[1], Fabio Pellegrini[2], Menglan Pang[2], Car De Moor[2], Changyu Shen[2], Vivek Charu[1] and ♦Lu Tian[1]. [1]Stanford University [2]Biogen

11:35 Floor Discussion.

Session 23INT50: Real World Evidence Research in Pharmaceutical Development

Room: 214
Organizer: Gang Li, Eisai Inc.
Chair: Gang Li, Eisai Inc..

10:20 Optimal Treatment Regimes: a Review and Empirical Comparison

Zhen Li[1], ♦Jie Chen[2], Eric Laber[3], Fang Liu[4] and Richard Baumgartner[4]. [1]Amazon [2]ECR Global [3]Duke [4]Merck

10:45 Floor Discussion.

Session 23INT51: Statistical Analysis of Streaming Data

Room: 215
Organizer: Ruijian Han, ruijian.han@polyu.edu.hk.
Chair: Huichen Zhu, The Chinese University of Hong Kong.

10:20 Online Inference with Debiased Stochastic Gradient Descent

♦*Ruijian Han[1], Lan Luo[2], Yuanyuan Lin[3] and Jian Huang[1].* [1]The Hong Kong Polytechnic University [2]University of Iowa [3]The Chinese University of Hong Kong

10:45 Inference in Heavy-Tailed Ar Models with Time Trends and Heteroscedastic Noises

Rui She. The Southwestern University of Finance and Economics

11:10 Irreversible Consumption Habit under Ambiguity: Singular Control and Optimal g-Stopping Time

Kyunghyun Park[1], ♦Kexin Chen[2] and Hoi Ying Wong[3]. [1]Nanyang Technological University [2]The Hong Kong Polytechnic University [3]The Chinese University of Hong Kong

11:35 Blocked Gibbs Sampler for Truncated Two-Parameter Poisson-Dirichlet Process

♦*Junyi Zhang[1] and Angelos Dassios[2].* [1]The Hong Kong Polytechnic University [2]London School of Economics

12:00 Floor Discussion.

July 8, 13:00-14:20

Session 23INTSP1: Special Invited Session

Room: LT1A
Organizer: ICSA Committee.
Chair: Xingqiu Zhao, The Hong Kong Polytechnic University.

13:00 Tba

Gang Li. UCLA

13:40 Semiparametric Predictive Inference for Failure Data using First-Hitting-Time Regression

♦*Mei-Ling Ting Lee[1] and George Whitmore[2].* [1]University of Maryland [2]McGill University

14:20 Floor Discussion.

July 8, 14:30-16:10

Session 23INT62: ML Meets Biostatistics: Theory and Practice

Room: LT1A
Organizer: Ben Dai, The Chinese University of Hong Kong.
Chair: Ben Dai, The Chinese University of Hong Kong.

14:30 Supervised Knowledge may Hurt Novel Class Discovery Performance

♦*Ziyun Li[1], Jona Otholt[1], Ben Dai[2], Di Hu[3], Christoph Meinel[1] and Haojin Yang[1].* [1]Hasso Plattner Institute [2]Chinese University of Hong Kong [3]Renmin University of China

14:55 De-Confounding Causal Inference using Latent Multiple-Mediator Pathways

♦*Yubai Yuan[1] and Annie Qu[2].* [1]The Pennsylvania State University [2]University of California, Irvine

15:20 Causal Inference in Transcriptome-Wide Association Studies with Invalid Instruments and Gwas Summary Data

♦*Haoran Xue, Xiaotong Shen and Wei Pan.* University of Minnesota

15:45 Knockofftrio: a Knockoff Framework for the Identification of Putative Causal Variants in Genome-Wide Association Studies with Trio Design

♦*Yi Yang*[1]*, Chen Wang*[2]*, Linxi Liu*[3]*, Joseph Buxbaum*[4]*, Zihuai He*[5] *and Iuliana Ionita-Laza*[2]. [1]City University of Hong Kong [2]Columbia University [3]University of Pittsburgh [4]Icahn School of Medicine at Mount Sinai [5]Stanford University

16:10 Floor Discussion.

Session 23INT6: Bridging Statistics and Computation in High-Dimensional Data Analysis
Room: LT1B
Organizer: Yuxin Chen, University of Pennsylvania.
Chair: Yuxin Chen, University of Pennsylvania.

14:30 Community Detection with Multiple Source of Information

♦*Zongming Ma and Sagnik Nandy.* University of Pennsylvania

14:55 Ranking Inferences Based on the Top Choice of Multiway Comparisons

Jianqing Fan[1]*, Zhipeng Lou*[1]*,* ♦*Weichen Wang*[2] *and Mengxin Yu*[1]. [1]Princeton University [2]The University of Hong Kong

15:20 using Svd for Topic Modeling

Tracy Ke. Harvard University

15:45 Approximate Message Passing from Random Initialization with Applications to Z2 Synchronization

♦*Gen Li, Wei Fan and Yuting Wei.* University of Pennsylvania

16:10 Floor Discussion.

Session 23INT52: False Discovery Rate Control and Replicability Analysis of High Throughput Experiments
Room: 201
Organizer: Hongyuan Cao, Florida State University.
Chair: Hongyuan Cao, Florida State University.

14:30 Statistical Assessment of Replicability: New Concepts and Methods

Xiaoquan Wen. University of Michigan

14:55 An Innovative Nonparametric Procedure to Assess Reproducibility Across High-Throughput Studies

♦*Wen Zhou*[1]*, Austin Ellingworth*[1]*, Debashis Ghosh*[2] *and Zhigen Zhao*[3]. [1]Colorado State University [2]University of Colorado Denver - Anschutz Medical Campus [3]Temple University

15:20 A New Fdr Controlling Procedure for Identifying Simultaneous Signals

Linsui Deng[1]*, Kejun He*[1] *and* ♦*Xianyang Zhang*[2]. [1]Renmin University of China [2]Texas A&M University

15:45 Assessing Reproducibility of High-Throughput Experiments in the Case of Missing Data

Roopali Singh[1]*, Feipeng Zhang*[2] *and* ♦*Qunhua Li*[1]. [1]Penn State University [2]Xi'an JiaoTong Universitty

16:10 Floor Discussion.

Session 23INT81: Casual Inference in Biomedical Applications
Room: 202
Organizer: Ying Zhang, University of Nebraska Medical Center.
Chair: Junyi Zhou, Amgen Inc.

14:30 The Synthetic Instrument: From Sparse Association to Sparse Causation

Dingke Tang, Dehan Kong and ♦*Linbo Wang.* University of Toronto

14:55 Decision Trees with Fused Leaves for Prostate Cancer Diagnosis

Xiaogang Su. University of Texas at El Paso

15:20 A Reference-Free r-Learner for Treatment Recommendation

♦*Junyi Zhou*[1]*, Ying Zhang*[2] *and Wanzhu Tu*[3]. [1]Amgen Inc [2]University of Nebraska Medical Center [3]Indiana University-School of Medicine and Fairbanks School of Public Health

15:45 Transporting Randomized Trial Results to Estimate Counterfactual Survival Functions in Target Populations

Zhiqiang Cao[1]*,* ♦*Youngjoo Cho*[2] *and Fan Li*[3]. [1]Shenzhen Technology University [2]Konkuk University [3]Yale University

16:10 Floor Discussion.

Session 23INT60: Stein's Method and Statistical Applications
Room: 203
Organizer: Xiao Fang, The Chinese University of Hong Kong.
Chair: Xiao Fang, The Chinese University of Hong Kong.

14:30 Cramér-Type Moderate Deviation for Quadratic Forms with a Fast Rate

Xiao Fang[1]*,* ♦*Song-Hao Liu*[2] *and Qi-Man Shao*[2]. [1]The Chinese University of Hong Kong [2]Southern University of Science and Technology

14:55 Bounds for the Asymptotic Distribution of the Likelihood Ratio

♦*Andreas Anastasiou*[1] *and Gesine Reinert*[2]. [1]University of Cyprus [2]University of Oxford

15:20 Bootstrap Test for Multi-Scale Lead-Lag Relationships in High-Frequency Data

Takaki Hayashi[1] *and* ♦*Yuta Koike*[2]. [1]Keio University [2]University of Tokyo

15:45 Cramér-Type Moderate Deviation for Sums of Local Dependent Random Variables

Song-Hao Liu and ♦Zhuo-Song Zhang. Southern University of Science and Technology

16:10 Floor Discussion.

Session 23INT37: Bayesian Methods on Latent Variable Models
Room: 209A
Organizer: Xinyuan Song, The Chinese University of Hong Kong.
Chair: Zhixiang Lin, The Chinese University of Hong Kong.

14:30 Bayesian Diagnostics of Hidden Markov Structural Equation Models with Missing Data

♦Jingheng Cai[1], Ming Ouyang[2], Kai Kang[1] and Xinyuan Song[3]. [1]Sun Yat-sen University [2]The Chinese University of Hong Kong [3]The Chinese University of Hong Kong,

14:55 Variational Bayesian Inference for Two-Part Latent Variable Model

Yemao Xia.

15:20 Latent Multiple Mediation Analysis with the Bayesian Lasso

Lijin Zhang[1] and ♦Junhao Pan[2]. [1]Graduate School of Education, Stanford University, U.S. [2]Department of Psychology, Sun Yat-sen University, Guangzhou, China

15:45 Joint Analysis of Mixed Types of Outcomes with Latent Variables

♦Deng Pan[1], Yingying Wei[2] and Xinyuan Song[2]. [1]School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China [2]Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong

16:10 Floor Discussion.

Session 23INT80: Recent Developments in Survival Analysis
Room: 209B
Organizer: Xingqiu Zhao, The Hong Kong Polytechnic University.
Chair: Xingqiu Zhao, The Hong Kong Polytechnic University.

14:30 Health Utility Survival for Randomized Clinical Trials: Extensions and Statistical Properties

Yangqing Deng[1], ♦Meiling Hao[2], John R. De Lmeida[3] and Wei Xu[4]. [1]Princess Margaret Cancer Centre, University Health Network, [2]University of International Business and Economics [3]Department of Otolaryngology— H&N Surgery, University Health Network [4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

14:55 Transfer Learning by Optimal Model Averaging for Censored Data

♦Baihua He[1] and Xinyu Zhang[2]. [1]University of Science and Technology of China [2]Chinese Academy of Sciences

15:20 Simultaneous Variable Selection and Estimation for Interval-Censored Failure Time Data with Ancillary Information

♦Mingyue Du[1] and Xingqiu Zhao[2]. [1]The Hong Kong Polytechnic University Shenzhen Research Institute [2]The Hong Kong Polytechnic University

15:45 Floor Discussion.

Session 23INT78: Recent Advances in Long-Run Variance Estimation in Time Series and Spatial Data
Room: 214
Organizer: Kin Wai Chan, The Chinese University of Hong Kong.
Chair: Cheuk Hin Cheng, The Chinese University of Hong Kong.

14:30 Variance Estimation of Spatial Autocorrelated Data under Non-Constant Mean

♦Di Su and Kin Wai Chan. The Chinese University of Hong Kong

14:55 Revamping Kernel-Based Long-Run Variance Estimation: a Converging Kernel Approach

♦Xu Liu and Kin Wai Chan. The Chinese University of Hong Kong

15:20 Reorder-Based Variance Estimator in Nonparametric Regression with Time Series Errors

♦Zerun Wang and Kin Wai Chan. The Chinese University of Hong Kong

15:45 Recursive Nonparametric Estimation: Principles, Methods and Applications

♦Man Fung Leung[1] and Kin Wai Chan[2]. [1]University of Illinois Urbana-Champaign [2]The Chinese University of Hong Kong

16:10 Floor Discussion.

July 8, 16:30-18:10

Session 23INT107: Recent Advances in Statistical Methods for Analyzing High-Dimensional Cancer and Disease Surveillance Data
Room: LT1A
Organizer: Yi Li, University of Michigan.
Chair: Subharup Guha, University of Florida.

16:30 High Dimensional Gaussian Graphical Regression Models with Covariates

Jingfei Zhang[1] and ♦Yi Li[2]. [1]Univ of Miami [2]Univ of Michigan

16:55 Approximate Bayesian Computation Estimation of Models for the Natural History of Breast Cancer, with Application to Data from a Milan Cohort Study

♦Marco Bonetti[1], Laura Bondi[2], Denitsa Grigorova[3] and Antonio Russo[4]. [1]Bocconi University, Milan, Italy [2]Cambridge University, Cambridge, UK [3]Sofia University, Sofia, Bulgaria [4]UOC Osservatorio Epidemiologico, ATS, Milan, Italy

17:20 New Statistical Method for Spatio-Temporal Surveillance of Infectious Diseases
Peihua Qiu. University of Florida

17:45 Floor Discussion.

Session 23INT42: Recent Advances and Applications of Survival Analysis in Biomedical Research
Room: LT1B
Organizer: Yi Xiong, University of Mantioba; Fred Hutchinson Cancer Center.
Chair: Yi Xiong, University of Mantioba; Fred Hutchinson Cancer Center.

16:30 Cox Proportional Hazards Regression with Interval Censored Outcome and Covariate
♦*Dongdong Li*[1], *Yue Song*[2], *Wenbin Lu*[3], *Huldrych Gunthard*[4], *Roger Kouyos*[4] and *Rui Wang*[1]. [1]Harvard Medical School [2]Harvard School of Public Health [3]North Carolina State University [4]University of Zurich

16:55 Recent Advances in Handling Time-to-Event Data with Internal Covariates
♦*Trevor Thomson, Joan Hu and Bohdan Nosyk.* Simon Fraser University

17:20 Deep Neural Network with a Smooth Monotonic Output Layer for Dynamic Risk Prediction
Zhiyang Zhou. University of Manitoba

17:45 Floor Discussion.

Session 23INT64: Statistical Methods in Data Integration and Synthesis
Room: 201
Organizer: Shouhao Zhou, Pennsylvania State University.
Chair: Shouhao Zhou, Pennsylvania State University.

16:30 Repro Samples Method for Uncertainty Quantification in Irregular Inference Problems and more
Minge Xie. Rutgers University

16:55 Meta-Analysis of Safety Data
Shouhao Zhou. Penn State University

17:20 A Unifying Dependent Combination Framework with Applications to Association Tests
♦*Xiufan Yu*[1], *Linjun Zhang*[2], *Arun Srinivasan*[3], *Lingzhou Xue*[4] and *Minge Xie*[2]. [1]University of Notre Dame [2]Rutgers University [3]GSK plc [4]Penn State University

17:45 Optimizing Information Borrowing for Bayesian Hierarchical Model in Subgroup Analysis
Xuetao Lu and ♦*J. Jack Lee.* University of Texas MD Anderson Cancer Center

18:10 Floor Discussion.

Session 23INT4: Recent Development on Analysis of Complex Time-to-Event Data
Room: 202
Organizer: Donglin Zeng, University of North Carolina.
Chair: Siming Zheng, The Chinese University of Hong Kong.

16:30 Feature Screening with Large Scale and High Dimensional Survival Data
Grace Yi[1], ♦*Wenqing He*[1] and *Raymond Carroll*[2]. [1]University of Western Ontario [2]Texas A&M University and University of Technology Sydney

16:55 Linearized Maximum Rank Correlation Estimation
Guohao Shen[1], *Kani Chen*[2], *Jian Huang*[1] and ♦*Yuanyuan Lin*[3]. [1]The Hong Kong Polytechnic University [2]Hong Kong University of Science and Technology [3]The Chinese University of Hong Kong

17:20 Efficient Estimation for the Accelerated Failure Time Model with Auxiliary Aggregated Information
♦*Huijuan Ma*[1], *Yukun Liu*[1], *Donglin Zeng*[2] and *Yong Zhou*[1]. [1]East China Normal University [2]University of North Carolina

17:45 Floor Discussion.

Session 23INT24: Recent Advances in Nonparametric Statistics and Novel Applications
Room: 203
Organizer: Yichuan Zhao, Georgia State University.
Chair: Yichuan Zhao, Georgia State University.

16:30 Variable Selection in Semiparametric Transformation Regression with Interval-Censored Competing Risks Data
Fatemeh Mahmoudi and ♦*Xuewen Lu.* University of Calgary

16:55 Optimal-k Sequence for Difference-Based Methods in Nonparametric Regression
Wenlin Dai[1], *Xingwei Tong*[2] and ♦*Tiejun Tong*[3]. [1]Renmin University of China [2]Beijing Normal University [3]Hong Kong Baptist University

17:20 Hypotheses Testing of Functional Principal Components
Zening Song[1], ♦*Lijian Yang*[2] and *Yuanyuan Zhang*[3]. [1]Nankai University [2]Tsinghua University [3]Soochow University

17:45 A Nearest Neighbor Method for Continuous Stochastic Optimization
♦*Seksan Kiatsupaibul*[1], *Pariyakorn Maneekul*[2] and *Zelda Zabinsky*[2]. [1]Chulalongkorn University [2]University of Washington

18:10 Floor Discussion.

Session 23INT39: Recent Developments on Complex Data Analysis
Room: 209A
Organizer: Xinyuan Song, The Chinese University of Hong Kong.
Chair: Xiangnan Feng, Fudan University.

16:30 Earthquake Parametric Insurance with Bayesian Spatial Quantile Regression
Jefferey Pai[1], ♦*Yunxian Li*[2], *Aijun Yang*[3] and *Chenxu Li*[4]. [1]University of Manitoba [2]University of Yunnan University of Finance and Economics [3]Nanjing Forest University [4]Yunnan University of Finance and Economics

16:55 Optimal Integrating Learning for Split Questionnaire Design Type Data

Cunjie Lin[1], Jingfu Peng[1], Yichen Qin[2], ♦Yang Li[1] and Yuhong Yang[3]. [1]Renmin University of China [2]University of Cincinnati [3]University of Minnesota

17:20 Robust Estimation and Test Based on Median-of-Means Method

Pengfei Liu. Jiangsu Normal University

17:45 Additive Hazards Model with Time-Varying Coefficients and Imaging Predictors

♦*Qi Yang[1], Chuchu Wang[2], Haijin He[3], Xiaoxiao Zhou[2] and Xinyuan Song[2].* [1]School of Management, Shandong University [2]The Chinese University of Hong Kong [3]Shenzhen University

18:10 Floor Discussion.

Session 23INT89: Challenges and Developments in Econometrics and Statistical Theories
Room: 209B
Organizer: Yuanyuan Lin, The Chinese University of Hong Kong.
Chair: Rongmao Zhang, Zhejiang University.

16:30 Spiked Eigenvalues of High-Dimensional Sample Autocovariance Matrices: Clt and Applications

Daning Bi[1], Xiao Han[2], Adam Nie[1] and ♦Yanrong Yang[1]. [1]Australian National University [2]University of Science and Technology of China

16:55 A General m-Estimation Theory in Semi-Supervised Framework

♦*Shanshan Song[1], Yuanyuan Lin[1] and Yong Zhou[2].* [1]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China [2]KLATASDS-MOE, School of Statistics and Academy of Statistics andInterdisciplinary Sciences, East China Normal University, Shanghai, China

17:20 Semi-Supervised Inference for Nonparametric Logistic Regression

♦*Tong Wang[1], Wenlu Tang[2], Yuanyuan Lin[1] and Wen Su[3].* [1]Department of Statistics, The Chinese University of Hong Kong [2]Department of Applied Mathematics, The Hong Kong Polytechnic University [3]Department of Statistics and Actuarial Science, The University of Hong Kong

17:45 Floor Discussion.

Session 23INT54: Statistical Methods in Health Research
Room: 214
Organizer: Yang Li, Indiana University School of Medicine and Health Data Science, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health.
Chair: Qingning Zhou, Mathematics & Statistics Department, University of North Carolina at Charlotte.

16:30 Linkage of Big Data of Electronic Medical Records in the Presence of Missing Data

♦*Xiaochun Li[1], Huiping Xu[1] and Shaun Grannis[2].* [1]Department of Biostatistics and Health Data Science, School of Medicine, Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana 46202, U.S.A. [2]Regenstrief Institute, Inc., Indianapolis, IN

16:55 Independent Screening for Nonparametric Additive Cox Model

Jiancheng Jiang.

17:20 Assessing Intra- And Inter-Method Agreement of Functional Data

Ye Yue[1], ♦Jeong Hoon Jang[2] and Amita Manatunga[1]. [1]Emory University [2]Yonsei University

17:45 Assessing Disparities in Americans' Exposure to Pcbs and Pbdes Based on Nhanes Pooled Biomonitoring Data

Yan Liu[1], ♦Dewei Wang[2], Li Li[1] and Dingsheng Li[1]. [1]University of Nevada, Reno [2]University of South Carolina

18:10 Floor Discussion.

Session 23INT63: Recent Advances in Computational Algorithms for Statistical Inference
Room: 215
Organizer: Zhou Fan, Yale University.
Chair: Zongming Ma, The Wharton School, University of Pennsylvania.

16:30 Sharp Analysis of Em for Learning Mixtures of Pairwise Differences

Abhishek Dhawan, ♦Cheng Mao and Ashwin Pananjady. Georgia Institute of Technology

16:55 Estimation of Leading Multi-Block Canonical Correlation Directions via ℓ_1-Norm Constrained Proximal Gradient Descent

Leying Guan. Yale University

17:20 Fundamental Limits of Spectral Clustering in Stochastic Block Models

Anderson Ye Zhang. University of Pennsylvania

17:45 Floor Discussion.

July 9, 9:00-10:00

Session 23INTKT3: Keynote Talk 3: Ji Zhu
Room: LT1A
Organizer: ICSA Committee.
Chair: Zhezhen Jin, Columbia University.

9:00 Statistical Inference on Latent Space Models for Network Data

Ji Zhu. University of Michigan

9:50 Floor Discussion.

July 9, 10:20-12:00

Session 23INT101: Statistical Machine Learning and Inference
Room: LT1A
Organizer: Jianqing Fan, Princeton University.
Chair: Jianqing Fan, Princeton University.

10:20 Simple-Rc: Group Network Inference with Non-Sharp Nulls and Weak Signals

Jianqing Fan[1], Yingying Fan[2], $\blacklozenge$Jinchi Lv[2] and Fan Yang[3]. [1]Princeton University [2]University of Southern California [3]Tsinghua University

10:45 A Non-Asymptotic Framework for the Approximate Message Passing Algorithm

Yuting Wei. University of Pennsylvania

11:10 Fairness-Adjusted Neyman-Pearson Classifiers

Ziqing Guo[1], Xin Tong[2] and $\blacklozenge$Lucy Xia[1]. [1]HKUST [2]USC

11:35 Online Tensor Learning: Computational and Statistical Trade-Offs, Adaptivity and Optimal Regret

Jian-Feng Cai, Jingyang Li and $\blacklozenge$Dong Xia. HKUST

12:00 Floor Discussion.

Session 23INT3: New Machine Learning and Semi-parametric Methods for Personalized Medical Decision Making
Room: LT1B
Organizer: Donglin Zeng, University of North Carolina.
Chair: Ruijian Han, The Hong Kong Polytechnic University.

10:20 Linear Discriminant Analysis with High-Dimensional Mixed Variables

Binyan Jiang. The Hong Kong Polytechnic University Shenzhen Research Institute

10:45 Recommending when to Treat: From Binary to Time-to-Intervention Decision

Li Hsu[1], $\blacklozenge$Yair Goldberg[2] and Yingye Zheng[1]. [1]Fred Hutchinson Cancer Research Center [2]Technion - Israel Institute of Technology

11:10 Matching-Based Learning for Decision Making using Electronic Health Records

Yuanjia Wang. Columbia University

11:35 Learning Optimal Group-Structured Individualized Treatment Rules

Haixu Ma, $\blacklozenge$Donglin Zeng and Yufeng Liu. University of North Carolina

12:00 Floor Discussion.

Session 23INT45: Recent Developments in Statistical Network Analysis
Room: 201
Organizer: Emma Jingfei Zhang, Emory University.
Chair: Emma Jingfei Zhang, Emory University.

10:20 Higher-Order Accurate Two-Sample Network Inference and Network Hashing

Meijia Shao[1], Dong Xia[2], $\blacklozenge$Yuan Zhang[1], Qiong Wu[3] and Shuo Chen[4]. [1]The Ohio State University [2]Hong Kong University of Science and Technology [3]University of Pennsylvania [4]University of Maryland, Baltimore

10:45 Limit Results for Distributed Estimation of Invariant Subspaces in Multiple Networks Inference and Pca

Runbing Zheng and $\blacklozenge$Minh Tang. North Carolina State University

11:10 Subsampling-Based Modified Bayesian Information Criterion for Large-Scale Stochastic Block Models

Jiayi Deng[1], $\blacklozenge$Danyang Huang[1], Xiangyu Chang[2] and Bo Zhang[1]. [1]Renmin University of China [2]Xi'an Jiaotong University

11:35 Learning Network Properties without Network Data – a Correlated Network Scale-Up Model

Ian Laga[1], Le Bao[2] and $\blacklozenge$Xiaoyue Niu[2]. [1]Montana State University [2]Penn State University

12:00 Floor Discussion.

Session 23INT43: Modern Machine Learning Approaches for Efficient Estimation and Sampling
Room: 202
Organizer: Junhui Wang, Chinese University of Hong Kong.
Chair: Junhui Wang, Chinese University of Hong Kong.

10:20 Bayesian Analysis for Functional Anova Model

Yongdai Kim. Seoul National University

10:45 Robust Estimation of Central Subspace under High-Dimensional and Elliptically-Contoured Design

$\blacklozenge$Jing Zeng[1] and Qing Mai[2]. [1]University of Science and Technology of China [2]Florida State University

11:10 Data-Adaptive Discriminative Feature Localization with Statistically Guaranteed Interpretation

$\blacklozenge$Ben Dai[1], Xiaotong Shen[2], Lin Yee Chen[2], Chunlin Li[2] and Wei Pan[2]. [1]The Chinese University of Hong Kong [2]University of Minnesota

11:35 Efficient Multimodal Sampling via Tempered Distribution Flow

$\blacklozenge$Yixuan Qiu[1] and Xiao Wang[2]. [1]Shanghai University of Finance and Economics [2]Purdue University

12:00 Floor Discussion.

Session 23INT25: Recent Development of Statistical Methods for Health Sciences
Room: 203
Organizer: Yanqing Sun, University of North Carolina at Charlotte, USA.
Chair: Yichuan Zhao, Georgia State University, USA.

10:20 Two-Sample Test and Support Recovery for Image Data

$\blacklozenge$Lianqiang Qu[1], Jian Huang, Liuquan Sun and Hongtu Zhu. [1]Central China Normal University

10:45 Deep Learning for Time-to-Event Predictions with Applications to Ehr Data

Xueying Wang[1], Jing Ning[2], Ruosha Li[1], Han Feng[3] and ♦Hulin Wu[1]. [1]University of Texas Health Science Center at Houston [2]University of Texas MD Anderson Cancer Center [3]Tulane University

11:10 Theoretical Properties of Oversampling and Subsampling for Imbalanced Classification

Jie Zhou.

11:35 Post-Episodic Reinforcement Learning Inference

Vasilis Syrgkanis[1] and ♦Ruohan Zhan[2]. [1]Stanford University [2]Hong Kong University of Science and Technology

12:00 Floor Discussion.

Session 23INT70: Quantile Regression with Complex Data

Room: 209A
Organizer: Huichen Zhu, The Chinese University of Hong Kong.
Chair: Guoyou Qin, Fudan University.

10:20 A Semiparametric Quantile Single-Index Model for Zero-Inflated and Overdispersed Outcomes

Tianying Wang. Tsinghua University

10:45 Fast Imputation Algorithms in Quantile Regression with Missing Covariates

Hao Cheng. National Academy of Innovation Strategy, China Association for Science and Technology

11:10 Censored Quantile Regression Forest

♦Huichen Zhu[1], Yifei Sun[2] and Ying Wei[2]. [1]The Chinses University of Hong Kong [2]Columbia University

11:35 Floor Discussion.

Session 23INT1: High-Dimensional Data Analysis

Room: 209B
Organizer: Mu Yue, Singapore University of Technology & Design (SUTD).
Chair: Mu Yue, Singapore University of Technology & Design (SUTD).

10:20 Ensemble Projection Pursuit for General Nonparametric Regression

Haoran Zhan, Mingke Zhang and ♦Yingcun Xia.

10:45 Inference on High-Dimensional Single-Index Models with Streaming Data

Dongxiao Han. Nankai University

11:10 High-Dimensional Covariance Matrices under Dynamic Volatility Models: Asymptotics and Shrinkage Estimation

Yi Ding[1] and ♦Xinghua Zheng[2]. [1]University of Macau [2]Hong Kong University of Science and Technology

11:35 Inference for Nonstationary Time Series with Varying Periodicity, a Smooth Trend and Covariate Effects

♦Ming-Yen Cheng[1], Shouxia Wang[1] and Lucy Xia[2]. [1]Hong Kong Baptist University [2]Hong Kong University of Science and Technology

12:00 Floor Discussion.

Session 23INT74: Modern Statistical and Machine Learning Modeling of Big Data

Room: 214
Organizer: Jingyi Zheng, Auburn University.
Chair: Jingyi Zheng, Auburn University.

10:20 Directly Deriving Parameters from Sdss Photometric Images

♦Fan Wu and Yude Bu.

10:45 Barycenter Estimation of Positive Semi-Definite Matrices with Bures-Wasserstein Distance

♦Jingyi Zheng[1], Huajun Huang[1], Yuyan Yi[1], Yuexin Li[1] and Shu-Chin Lin[2]. [1]Auburn University [2]National Health Research Institutes, Taiwan

11:10 Simultaneous Identification of Brain Functional Differential Network and Gene Regulatory Pathways

Hao Chen[1], Yong He[2] and ♦Jiadong Ji[2]. [1]Peking University [2]Shandong University

11:35 Fine-Mapping Causal Variants using Summary Gwas Statistics with Heritability-Induced Dirichlet Decomposition Prior

Xiang Li and ♦Yan Dora Zhang. The University of Hong Kong

12:00 Floor Discussion.

Session 23INT75: Recent Advances in Integrative Analysis of Multi-Omics Data

Room: 215
Organizer: Xuelin Huang, The University of Texas MD Anderson Cancer Center.
Chair: Kai Zhao, The Chinese University of Hong Kong.

10:20 Leveraging Genomic Regulations to Inform Genetic Association Analysis with Complex Traits

Li Hsu.

10:45 Discrete Representation Learning for Single-Cell Multi-Omics Data

Xuejian Cui[1], Shengquan Chen[2] and ♦Rui Jiang[1]. [1]Tsinghua University [2]Nankai University

11:10 Bayesian Inference for Non-Invasive Preimplantation Genetic Testing

♦Hao Ge, Ruiqi Zhang, Lei Huang and Xiaoliang Xie. Peking University

11:35 Statistical Methods for Mediation Analysis with High-Dimensional Omics Mediators

♦Peng Wei[1], Sunyi Chi[1], Zhichao Xu[1], Tianzhong Yang[2], Chunlin Li[2] and Xuelin Huang[1]. [1]The University of Texas MD Anderson Cancer Center [2]University of Minnesota

12:00 Floor Discussion.

Session 23INTSP3: Junior Researcher Award Session
Room: LT4
Organizer: Hongbin Fang, Georgetown University.
Chair: Hongbin Fang, Georgetown University.

10:20 Multi-State Model and Structural Selection for the Analysis of Depressive Symptom Dynamics in Middle-Aged and Older Adults

Chuoxin Ma.   BNU-HKBU United International College

10:45 Partial Quantile Tensor Regression

Dayu Sun.   Indiana University

11:10 Desiderata for Representation Learning: a Causal Perspective

Yixin Wang.   University of Michigan

11:35 Learning Network-Structured Dependence from Non-Stationary Multivariate Point Process Data

Muhong Gao.   Chinese Academy of Science

12:00 Floor Discussion.

July 9, 13:00-14:40

Session 23INT55: Recent Advances of High-Dimensional Models and Time Series Models
Room: LT1A
Organizer: Quefeng Li, UNC at Chapel Hill.
Chair: Quefeng Li, UNC at Chapel Hill.

13:00 Optimal and Safe Estimation for High-Dimensional Semi-Supervised Learning

Yang Ning.   Cornell University

13:25 Cross-Covariance-Based Estimator of Autocorrelation Operator for Arh(1) Process with Contamination

♦*Zhao Chen*[1], *Liang Hu*[2] *and Yuwei Zhao*[1].   [1]Fudan University [2]Chinese Academy of Sciences

13:50 Inference for High-Dimensional Linear Models with Locally Stationary Error Processes

Jiaqi Xia, Yu Chen and ♦*Xiao Guo*.   University of Science and Technology of China

14:15 Normal Mixture Quasi-Maximum Likelihood Estimation of Double Autoregressive Models

Zhao Chen[1], *Min Chen*[2] *and* ♦*Christina Dan Wang*[3].   [1]Fudan University [2]AMSS [3]NYU Shanghai

14:40 Integrative Analysis of Gaussian Graphical Models

Shuangge Ma.   Yale University

Floor Discussion.

Session 23INT21: Recent Developments for Dependent Data with Complex Structure
Room: LT1B
Organizer: Guanyu Hu, University of Missouri – Columbia.
Chair: Yang Ni, Texas A& M University.

13:00 A Variational Bayesian Approach to Identifying Whole-Brain Directed Networks with Fmri Data

Yaotian Wang[1], *Guofen Yan*[2], *Xiaofeng Wang*[3], *Shuoran Li*[1], *Lingyi Peng*[1], *Dana Tudorascu*[1] *and* ♦*Tingting Zhang*[1].   [1]University of Pittsburgh [2]University of Virginia [3]Cleveland Clinic

13:25 Bayesian Image Mediation Analysis

Yuliang Xu and ♦*Jian Kang*.   University of Michigan

13:50 Precision Education: a Bayesian Nonparametric Approach for Handling Item and Examinee Heterogeneity in Assessment Data

Tianyu Pan[1], *Weining Shen*[1], *Clintin Stober*[2] *and* ♦*Guanyu Hu*[2].   [1]University of California Irvine [2]University of Missouri Columbia

14:15 High-Dimensional Response Growth Curve Modeling for Longitudinal Neuroimaging Analysis

♦*Lu Wang*[1], *Xiang Lyu*[2], *Zhengwu Zhang*[3] *and Lexin Li*[2].   [1]Central South University [2]University of California at Berkeley [3]University of North Carolina at Chapel Hill

14:40 Floor Discussion.

Session 23INT83: Statistical Learnning on Complex Data
Room: 201
Organizer: Ting Li, The Hong Kong Polytechnic University.
Chair: Binyan Jiang, The Hong Kong Polytechnic University.

13:00 Deep Kronecker Network

♦*Long Feng*[1] *and Guang Yang*[2].   [1]University of Hong Kong [2]City University of HOng Kong

13:25 Fpls-Dc: Functional Partial Least Squares Through Distance Covariance for Imaging Genomics

♦*Wenliang Pan*[1], *Chuang Li*[2], *Yue Shan*[3], *Tengfei Li*[3], *Yun Li*[3] *and Hongtu Zhu*[3].   [1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences [2]Sun Yat-sen university [3]University of North Carolina at Chapel Hill

13:50 Ball Impurity: Measuring Heterogeneity in General Metric Spaces

Ting Li.   The Hong Kong Polytechnic Universtiy

14:15 Floor Discussion.

Session 23INT68: Statistical Design and Analysis of Reliability and Survival Data
Room: 202
Organizer: Hon Keung Tony Ng, Bentley University.
Chair: Hon Keung Tony Ng, Bentley University.

13:00 Kaplan-Meier Type Precedence Test Based on Ranked Set Progressively Type-Ii Censored Data

Chang Cui, ♦*Tao Li, Jiaqi Men and Yijia Zheng*.   Shanghai University of Finance and Economics

13:25 Minimax Designs for Accelerated Life Tests

♦*I-Chen Lee*[1], *Ray-Bing Chen*[1] *and Weng Kee Wong*[2].   [1]National Cheng Kung University [2]University of California, Los Angeles

13:50 Reliability Estimation for One-Shot Devices with Correlated Components

Man Ho Ling. The Education University of Hong Kong

14:15 Floor Discussion.

Session 23INT72: Recent Advancements in Statistical Methods for Complex Lifetime Data

Room: 203
Organizer: Jing Ning, The University of Texas MD Anderson Cancer Center.
Chair: Wen Li, The University of Texas McGovern Medical School at Houston.

13:00 Evaluation of the Natural History of Disease by Combining Incident and Prevalent Cohorts: Application to the Nun Study

♦*Daewoo Pak[1], Jing Ning[2], Richard Kryscio[3] and Yu Shen[2]*. [1]Yonsei University [2]The University of Texas MD Anderson Cancer Center [3]University of Kentucky

13:25 Interval-Censored Linear Rank Regression

♦*Sangbum Choi[1], Taehwa Choi[2] and Wenbin Lu[3]*. [1]Department of Statistics, Korea University [2]Department of Biostatistics and Bioinformatics, Duke University [3]Department of Statistics, North Carolina State University

13:50 Quantile Association Regression on Bivariate Survival Data

Ling-Wan Chen[1], ♦*Yu Cheng[2], Ying Ding[2] and Ruosha Li[3]*. [1]FDA [2]University of Pittsburgh [3]UT Health

14:15 Floor Discussion.

Session 23INT77: Recent Advances in Nonparametric Methods in Time Series and Econometrics

Room: 209A
Organizer: Kin Wai Chan, The Chinese University of Hong Kong.
Chair: Man Fung Leung, University of Illinois Urbana-Champaign.

13:00 Mean Stationarity Test in Time Series: a Signal Variance-Based Approach

♦*Hon Kiu To and Kin Wai Chan*. The Chinese University of Hong Kong

13:25 Optimally Jittered Jump Test for High-Frequency Data

Cheuk Hin Cheng, Hon Kiu To, Kai Pan Chu and ♦*Ting Tin Ma*. Department of Statistics, CUHK

13:50 A Non-Parametric Approach for Causal Inference in Time Series

♦*Kai Pan Chu and Kin Wai Chan*. Department of Statistics, The Chinese University of Hong Kong

14:15 General Framework for Self-Normalized Multiple-Change-Point Tests

♦*Cheuk Hin Cheng and Kin Wai Chan*.

14:40 Floor Discussion.

Session 23INT79: Semiparametric Inference for Complex Data

Room: 209B
Organizer: Xingqiu Zhao, The Hong Kong Polytechnic University.
Chair: Xingqiu Zhao, The Hong Kong Polytechnic University.

13:00 Multiple Descent in Random Feature Regression

Xuran Meng[1], Jianfeng Yao[2] and ♦*Yuan Cao[1]*. [1]The University of Hong Kong [2]The Chinese University of Hong Kong (Shenzhen)

13:25 Phase-Type Sieve

Zhisheng Ye. National University of Singapore

13:50 Temporal Heterogeneity Learning for Functional Panel Quantile Regressions

♦*Jiaqi Men[1], Jinhong You[1] and Hua Liu[2]*. [1]Shanghai University of Finance and Economics [2]Xi'an Jiaotong University

14:15 Floor Discussion.

Session 23INT87: Network Structure and Structural Change-Point Estimation

Room: 214
Organizer: Min Xu, Department of Statistics, Rutgers University.
Chair: Min Xu, Department of Statistics, Rutgers University.

13:00 Sparse Change Detection in High-Dimensional Linear Regression

Fengnan Gao[1] and ♦*Tengyao Wang[2]*. [1]Fudan University [2]LSE

13:25 Changepoint Detection in Preferential Attachment Networks

Daniel Cirkovic[1], ♦*Tiandong Wang[2] and Xianyang Zhang[1]*. [1]Texas A&M University [2]Fudan University

13:50 Community Detection in Sparse Latent Space Models

♦*Fengnan Gao[1], Hongsong Yuan and Zongming Ma*. [1]Fudan University

14:15 Root and Community Inference on Markovian Models of Networks

Min Xu. Rutgers University

14:40 Floor Discussion.

Session 23INTSP2: Special Memorial Session to Celebrate Life of Professor Tze Leung Lai

Room: LT4
Organizer: TBA.
Chair: TBA.

13:00 Tba

Gang Li. UCLA

13:25 Advancing Sequential Importance Sampling: a Tribute to Professor Tze Leung Lai

Yuguo Chen. University of Illinois Urbana-Champaign

13:50 Innovations in Clinical Trial Methodology for Precision Medicine: a Tribute to Professor Tze Leung Lai

    ♦*Ying Lu and Lu Tian*.   Stanford University

14:15 Professor Lai's Contributions to Sequential Experimentation

    Zhiliang Ying.   Columbia University

14:40 Professor Lai's Contributions and Influence on Statistical Research in Taiwan

    ♦*Ching-Kang Ing[1], I-Ping Tu[2] and Chao A. Hsiung[3]*. [1]National Tsing Hua University [2]Academia Sinica, Taiwan [3]National Health Research Institutes, Taiwan

    Floor Discussion.

Session 23INT109: Complex Data Analysis
Room: 215
Organizer: Xueqin Wang, University of Science and Technology of China.
Chair: Canhong Wen, University of Science and Technology of China.

13:00 Covariate Balancing with Measurement Error

    Ying Yan.   Sun Yat-sen University

13:25 Estimation and Model Selection for Nonparametric Function-on-Function Regression

    ♦*Zhanfeng Wang[1], Hao Dong[2], Ping Ma[3] and Yuedong Wang[2]*.   [1]University of Science and Technology of China [2]University of California, Santa Barbara [3]University of Georgia

13:50 Deep Image-on-Scalar Regression Model with Hidden Confounders

    ♦*Xiaohe Chen, Lintao Tang, Rongjie Liu and Chao Huang*.   Florida State University

14:15 Floor Discussion.

July 9, 14:50-16:30

Session 23INT91: Recent Developments in Biostatistics with their Applications
Room: LT1A
Organizer: Judy Zhong, Department of Population Health, NYU Grossman School of Medicine.
Chair: Judy Zhong, Department of Population Health, NYU Grossman School of Medicine.

14:50 Recent Developments in Biostatistics with their Applications

    Chen Lyu.   Department of Population Health, NYU Grossman School of Medicine

15:15 Accurate Estimation of Breakpoints in Piecewise Linear Mixed-Effects Models with Application to Longitudinal Ophthalmic Studies

    Jiyuan Hu.   NYU Grossman School of Medicine

15:40 Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-Omics Data, and Predicting Disease Risk

    ♦*Chan Wang, Leopoldo Segal, Jiyuan Hu, Boyan Zhou, Richard Hayes, Jiyoung Ahn and Huilin Li*.   New York University Grossman School of Medicine

15:05 Floor Discussion.

Session 23INT92: Incorporating External Data in Superiority and Non-Inferiority Clinical Trials: Bayesian Nonparametric vs Parametric Models
Room: LT1B
Organizer: Jun Yin, Mayo Clinic.
Chair: Jun Yin, Mayo Clinic.

14:50 A Bayesian Nonparametric Model for External Data with Application to Clinical Trials

    ♦*Dehua Bi and Yuan Ji*.   The University of Chicago

15:15 A Bayesian Parametric Model to Incorporate Real-World Evidence in Pragmatic Trials

    ♦*Jun Yin, Peter Noseworthy and Xiaoxi Yao*.   Mayo Clinic

15:40 Floor Discussion.

Session 23INT22: Advances in Statistical Genetics and Genomics
Room: 201
Organizer: Jin Liu, Chinese University of Hong Kong, Shenzhen.
Chair: Jin Liu, Chinese University of Hong Kong, Shenzhen.

14:50 Funcode: Scoring Cross-Species Functional Conservation of Dna Elements using Encode Data

    Weixiang Fang[1], Chaoran Chen[2], Boyang Zhang[1], Yi Wang[1], Ruzhang Zhao[1], Weiqiang Zhou[1] and ♦*Hongkai Ji[1]*.   [1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health [2]Department of Biomedical Engineering, Johns Hopkins University

15:15 Probabilistic Cell/Domain-Type Assignment of Spatial Transcriptomics Data with Spatialanno

    Xingjie Shi.   East China Normal University

15:40 Evaluation of Epitranscriptome-Wide N6-Methyladenosine Di erential Analysis Methods

    Daoyu Duan[1], Wen Tang[1], Runshu Wang[2], ♦*Zhenxing Guo[3] and Hao Feng[1]*.   [1]Department of Population and Quantitative Health Sciences, Case Western Reserve University [2]Department of Biostatistics, University of Michigan [3]School of Data Science, The Chinese University of Hong Kong - Shenzhen

15:05 Rna Velocity Estimation with Stochastic Differential Equations

    ♦*Xu Liao[1], Lican Kang[1], Xiaoran Chai[1], Yuling Jiao[2] and Jin Liu[3]*.   [1]National University of Singapore [2]Wuhan University [3]The Chinese University of Hong Kong, Shenzhen

16:30 Floor Discussion.

Session 23INT9: Bayesian Spatial Analysis: Theory, Method, and Application
Room: 202
Organizer: Guanyu Hu, University of Missouri - Columbia.
Chair: Guanyu Hu, University of Missouri - Columbia.

14:50 Bayesian Fixed-Domain Asymptotics for Covariance Parameters in Spatial Gaussian Process Regression Models
♦*Cheng Li*[1], *Saifei Sun*[1] *and Yichen Zhu*[2]. [1]National University of Singapore [2]Duke University

15:15 Bayesian Modeling of Spatially Resolved Transcriptomics Data
Qiwei Li. The University of Texas at Dallas

15:40 Characterizing the Extremal Dependence in Spatial Analysis of 2021 Pacific Northwest Heatwave
♦*Likun Zhang*[1], *Mark Risser*[2], *Michael Wehner*[2] *and Travis O'brien*[3]. [1]University of Missouri [2]Lawrence Berkeley National Laboratory [3]Indiana University

15:05 Uncertainty Quantification for Object Boundaries Extracted from Spatial Point Patterns
Jue Wang[1], *Vinay Kashyap*[2], ♦*Thomas Lee*[1], *David Van Dyk*[3] *and Andreas Zezas*[4]. [1]UC Davis, USA [2]Center for Astrophysics | Harvard & Smithsonian, USA [3]Imperial College London, UK [4]deUniversity of Crete, Greece

16:30 Floor Discussion.

Session 23INT73: Challenges and Advances in Risk Assessment and Prediction
Room: 203
Organizer: Jing Ning, The University of Texas MD Anderson Cancer Center.
Chair: Daewoo Pak, Division of Data Science, Yonsei University.

14:50 Conditional Concordance-Assisted Learning for Combining Biomarkers for Population Screening
♦*Wen Li*[1], *Ruosha Li*[2], *Qingxiang Yan*[3], *Ziding Feng*[4] *and Jing Ning*[5]. [1]The University of Texas McGovern Medical School at Houston, TX, USA [2]The University of Texas School of Public Health, TX, USA [3]F. Hoffmann-La Roche Ltd., ON, Canada [4]Fred Hutchinson Cancer Research Center, WA, USA [5]The University of Texas MD Anderson Cancer Center, TX, USA

15:15 Semiparametric Isotonic Regression Model and Estimation for Group Testing Data
Ao Yuan[1], ♦*Jin Piao*[2], *Jing Ning*[3] *and Jing Qin*[4]. [1]Georgetown University [2]The University of Southern California [3]The University of Texas MD Anderson Cancer Center [4]National Institute of Allergy and Infectious Diseases

15:40 Analysis of Survival Data with Cure Fraction and Variable Selection: a Pseudo-Observations Approach
Chien-Lin Su[1], ♦*Sy Han Chiou*[2], *Feng-Chang Lin*[3] *and Robert Platt*[1]. [1]McGill University [2]University of Texas at Dallas [3]University of North Carolina

15:05 Floor Discussion.

Session 23INT93: Showcase of the Power of Statistics in Observational Studies for Precision Health
Room: 209A
Organizer: Lu Wang, University of Michigan.
Chair: Zhenke Wu, University of Michigan.

14:50 Statistical Opportunities in Analyzing Real-World Interventional Mobile Health Data
Zhenke Wu. University of Michigan, Ann Arbor

15:15 Constructing Time-Invariant Dynamic Surveillance Rules for Optimal Monitoring Schedules
Yingqi Zhao.

15:40 Reinforcement Learning for Estimating Optimal Dynamic Treatment Rules
Weijie Liang and ♦*Jinzhu Jia*. Peking University, China

15:05 Identify Sensitive Biomarkers of Alzheimer's Disease with Longitudinal Block-Wise Missing Data using Multiple Imputations Across Different Sources
♦*Zhongzhe Ouyang and Lu Wang*. University of Michigan, Ann Arbor

16:30 Floor Discussion.

Session 23INT97: Recent Developments on the Analysis of Censored Data
Room: 209B
Organizer: Kin Yau Wong, The Hong Kong Polytechnic University.
Chair: Kin Yau Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University.

14:50 On Interquantile Smoothness of Censored Quantile Regression with Induced Smoothing (Cqris)
♦*Zexi Cai*[1] *and Tony Sit*[2]. [1]Columbia University [2]The Chinese University of Hong Kong

15:15 Distributed Censored Quantile Regression
♦*Tony Sit*[1] *and Kelly Xing*[2]. [1]CUHK [2]Michigan State University

15:40 Semiparametric Regression Analysis of Doubly-Censored Data with Applications to Incubation Period Estimation
♦*Kin Yau Wong*[1], *Qingning Zhou*[2] *and Tao Hu*[3]. [1]The Hong Kong Polytechnic University [2]The University of North Carolina at Charlotte [3]Capital Normal University

15:05 A Semiparametric Joint Model for Cluster Size and Sub-unit-specific Interval-censored Outcomes
♦*Chun Yin Lee*[1], *Kin Yau Wong*[1], *Kwok Fai Lam*[2] *and Dipankar Bandyopadhyay*[3]. [1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong [2]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, Hong Kong [3]Department of Biostatistics, Virginia Commonwealth University, Virginia, USA

16:30 Floor Discussion.

Session 23INT99: Recent Development of Tensor Time Series
Room: 214
Organizer: Xinbing Kong, Nanjing Audit University.
Chair: Ting Li, The Hong Kong Polytechnic University.

14:50 Optimal Subsampling Bootstrap for Massive Data
♦*Yingying Ma*[1], *Chenlei Leng*[2] *and Hansheng Wang*[3]. [1]Beihang University [2]University of Warwick [3]Peking University

15:15 Huber Principal Component Analysis for Large-Dimensional Factor Model
♦*Yong He*[1], *Lingxiao Li*[1], *Dong Liu*[2] *and Wen-Xin Zhou*[3]. [1]Shandong University [2]Shanghai University of Finance and Economics [3]University of California, San Diego

15:40 Online Change-Point Detection for Matrix-Valued Time Series with Latent Two-Way Factor Structure
Yong He, Xinbing Kong, Lorenzo Trapani and ♦*Long Yu*.

15:05 Floor Discussion.

Session 23INT102: Statistical Inference on Complex/Compositional Data and Biostatistics
Room: 215
Organizer: Niansheng Tang, Yunnan University.
Chair: Niansheng Tang, Yunnan University.

14:50 Fdr Control for Linear log-Contrast Models with High-Dimensional Compositional Covariates
♦*Gaorong Li, Panxu Yuan and Changhan Jin*. Beijing Normal University

15:15 Renewable Quantile Regression with Heterogeneous Streaming Datasets
♦*Xuerong Chen and Senlin Yuan*.

15:40 Augmented Two-Step Estimating Equations with Nuisance Functionals and Complex Survey Data
♦*Puying Zhao*[1] *and Changbao Wu*[2]. [1]Yunnan University [2]University of Waterloo

15:05 Floor Discussion.

July 9, 16:50-18:30

Session 23INT29: Recent Advances on Interplay of Statistics and Optimization
Room: LT1A
Organizer: Anderson Ye Zhang, University of Pennsylvania.
Chair: Anderson Ye Zhang, University of Pennsylvania.

16:50 Self-Regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models
Yury Polyanskiy[1] *and* ♦*Yihong Wu*[2]. [1]MIT [2]Yale

17:15 Random Graph Matching at Otter's Threshold via Counting Chandeliers
Cheng Mao[1], *Yihong Wu*[2], ♦*Jiaming Xu*[3] *and Sophie Yu*[3]. [1]Georgia Institute of Technology [2]Yale University [3]Duke University

17:40 Snr Estimation under High-Dimensional Linear Models
Xiaohan Hu and ♦*Xiaodong Li*. UC Davis

17:05 Floor Discussion.

Session 23INT104: Innovative Designs and Analysis Methods for Clinical Trials and Complex Data
Room: LT1B
Organizer: Chunjie Wang, Changchun University of Technology.
Chair: Chunjie Wang, Changchun University of Technology.

16:50 Least Squares Support Vector Regression for Complex Censored Data
Xinrui Liu, Xiaogang Dong, Le Zhang, Jia Chen and ♦*Chunjie Wang*. Changchun University of Technology

17:15 Estimating Optimal Individual Treatment Regimes in Semi-Supervised Framework
♦*Mengjiao Peng and Yong Zhou*.

17:40 Floor Discussion.

Session 23INT59: Recent Advances in Biomedical Data Science
Room: 201
Organizer: Hongyu Zhao, Yale University.
Chair: Seyoung Park, Sungkyunkwan University.

16:50 Do We Need the Target-Decoy Strategy in Peptide Identification?
Sheng Lian[1], *Juntao Zhao*[1], *Zhen Zhang*[2], *Xiaodan Fan*[2], *Ning Li*[1] *and* ♦*Weichuan Yu*[1]. [1]HKUST [2]CUHK

17:15 Functional Adaptive Double-Sparsity Estimator for Functional Linear Regression with Applications in Kinect Sensor Analysis and Automated Elderly Health Assessments
Xinyue Li. City University of Hong Kong

17:40 A Multi-Use Graph Neural Network Framework for Single-Cell Multi-Omics Data
♦*Peifeng Ruan*[1] *and Hongyu Zhao*[2]. [1]UT Southwestern Medical Center [2]Yale University

17:05 Floor Discussion.

Session 23INT105: Recent Developments on Variable Selection and Regression Analysis with Censored Data
Room: 202
Organizer: Peijie Wang, Jilin University.
Chair: Peijie Wang, Jilin University.

16:50 Debiased Average Distance Correlation Screening of Massive Interval Censored Data under Orthogonal Subsampling
Huiqiong Li. Yunnan University

17:15 Bel and Bayesian Variables Selection for Partially Linear Models Based on Right-Censored Data
♦*Chunjing Li*[1] *and Xiaogang Dong*. [1]lichunjing@ccut.edu.cn

17:40 Simultaneous Variable Selection and Estimation for Joint Models of Longitudinal and Failure Time Data with Interval Censoring
Fengting Yi. Yunnan University

17:05 Semiparametric Regression Analysis of Length Biased Interval-Censored Data under the Mixture Cured Model
♦*Peijie Wang, Cunjin Zhao and Jianguo Sun*. Jilin University

18:30 Floor Discussion.

Session 23INT106: Analyzing Big and Complex Data using Modern Machine Learning Techniques
Room: 203
Organizer: Jiwei Zhao, University of Wisconsin-Madison.
Chair: Jiwei Zhao, University of Wisconsin-Madison.

16:50 Fragmgan: Generative Adversarial Nets for Fragmentary Data Imputation and Prediction
♦*Fang Fang and Shenliao Bao*. East China Normal University

17:15 Statistical Methods for Allele-Specific Expression Analysis using Single-Cell Rna-Seq Data
Rui Xiao. University of Pennsylvania

17:40 Robust Multiple Testing under High Dimensioanl Factor Model
Xinxin Yang[1] *and* ♦*Lilun Du*[2]. [1]INNO Asset Management [2]City University of Hong Kong

17:05 Learning Individualized Minimal Clinically Important Difference (Imcid) from High-Dimensional Data
Jiwei Zhao. University of Wisconsin Madison

18:30 Floor Discussion.

Session 23INT108: High-Dimensional Statistical Inference
Room: 209A
Organizer: Shurong Zheng, Northeast Normal University.
Chair: Shurong Zheng, Northeast Normal University.

16:50 Nonparametric Granger Causality Test Based on Adaptive Dimension Reducing Technique
♦*Yongchang Hui, Hong Jiang and Yi Liu*. Xi'an Jiaotong University

17:15 Change Point Inference in the High-Dimensional Correlation Matrix
♦*Zhaoyuan Li and Jie Gao*. CUHK-SZ

17:40 Separable Sample Covariance Matrices under Elliptical Populations with Applications
Huiqin Li[1], *Guangming Pan*[2], ♦*Yanqing Yin*[1] *and Wang Zhou*[3]. [1]Chongqing University [2]Nanyang Technological University [3]National University of Singapore

17:05 Adaptive Tests for Bandedness of High-Dimensional Covariance Matrices
Xiaoyi Wang. Beijing Normal University

18:30 Floor Discussion.

Abstracts

Session 23INTKT1: Keynote Talk 1: Songxi Chen

Multi-Level Thresholding for Detecting Rare and Faint Signals

Songxi Chen

Peking University

The work considered in this talk was largely motivated by Professor Hall works on testing for high dimensional means. I will summarize latest research on detection of rare and faint signals in the mean and the covariance matrices, the two basic summary measures of distributions, by formulating multi-level thresholding tests (MTT). The detection boundary and the minimax properties of the MTTs are presented. The MTTs are shown to be powerful in detecting sparse and weak signals in thigh dimensional mean and covariances, leading to attractive detection boundary and attain the optimal minimax rate in the signal strength under different regimes of high dimensionality and the sparsity of the signals.

Session 23INT90: Recent Development of Design and Analysis in Oncology Clinical Trials

How Should We Compare Survival Outcomes with Nonproportional Hazards?

♦*Rick Chappell and Mitchell Paukner*

University of Wisconsin Madison

Textbooks describing how to analyze time-to-event outcomes in clinical trials tend to list a limited range of topics. Differences are often quantified using hazard ratios from the Cox model and its associated score, the log-rank test. Weighted rank tests may be presented, along with comparisons of landmarks and quantiles. All these have their disadvantages in terms of interpretation, convenience, and/or power. Furthermore, cancer immunotherapy and many other treatments in a variety of diseases can have delayed effects causing pairs of curves to diverge after months or years of followup. In such cases the log-rank test's power will be low and the associated hazard ratio estimate uninterpretable. Rank tests with increasing weights suffer from the paradoxical property of rewarding early failures. I will discuss various available alternatives including some which are quite new.

Immune-Oncology Agents: Endpoints and Designs

Hao Wang

Johns Hopkins University School of Medicine

As our understanding of the immune system and how it and cancers interact have led to huge advances in oncology. The discovery and development of immune checkpoint inhibition and drugs that target the tumor's effect on checkpoints have revolutionized the treatment of multiple cancers, particularly melanoma and non-small cell lung cancer. At the same time, researchers are developing vaccines to treat cancers. Observations of difference in behaviors of tumors after patients receive these agents have led researchers to reconsider traditional study endpoints and develop new ones. These innovations have also led to new designs. In this talk, we discuss several of the endpoints for

evaluating effect and effectiveness of immunotherapies in oncology. We look at endpoints that seek to elucidate immune responses in patients, as well as clinical endpoints. We point out several concerns related to the endpoints and consider some ways one might address these concerns.

On Statistical Inference of Multiple Competing Risks in Comparative Clinical Trials

Jiyang Wen, Mei-Cheng Wang and ♦*Chen Hu*

Johns Hopkins University

Competing risks data are commonly encountered in comparative clinical trials. Ignoring competing risks in survival analysis leads to biased risk estimates and improper conclusions. Often, one of the competing events is of primary interest and the rest competing events are handled as nuisances. These approaches can be inadequate when multiple competing events have important clinical interpretations and thus of equal interest. For example, in hospitalized critical care treatment trials, the outcomes are either death or discharge from hospital, which have completely different clinical implications and are of equal interest. In oncology trials, while composite endpoints, such as disease-free survival, are used frequently, it is often concerned that novel interventions do not necessarily impact all components of a composite endpoint equally. We develop nonparametric estimation and simultaneous inferential methods for multiple cumulative incidence functions (CIFs) and corresponding restricted mean times. Based on Monte Carlo simulations and a data analysis of a completed clinical trial, we demonstrate that the proposed method provides global insights of the treatment effects across multiple endpoints.

Session 23INT30: Statistical and Deep Learning for Survival Data

Deep Generative Estimation of Conditional Survival Function

Xingyu Zhou[1], Wen Su[2], Changyu Liu[3], Yuling Jiao[4], Xingqiu Zhao[5] and ♦*Jian Huang[5]*

[1]Dow Inc.

[2]The University of Hong Kong

[3]The Chinese University of Hong Kong

[4]Wuhan University

[5]The Hong Kong Polytechnic University

We propose a deep generative approach to nonparametric estimation of conditional survival and hazard functions with right-censored data. The key idea of the proposed method is to first learn a conditional generator for the joint conditional distribution of the observed time and censoring indicator given the covariates, and then construct the Kaplan-Meier and Nelson-Aalen estimators based on this conditional generator for the conditional hazard and survival functions. Our method combines ideas from the recently developed deep generative learning approach and classical nonparametric estimation in survival analysis. We analyze the convergence properties of the proposed method and establish the consistency of the generative

nonparametric estimators of the conditional survival and hazard functions. Our numerical experiments validate the proposed method and demonstrate its superior performance in a range of simulated models. We also illustrate the applications of the proposed method in constructing prediction intervals for survival times with two data examples.

Deep Learning for the Partially Linear Cox Model

♦Qixian Zhong[1], Jonas Mueller[2] and Jane-Ling Wang[3]
[1]Xiamen University
[2]Cleanlab
[3]UC Davis

While deep learning approaches to survival data have demonstrated empirical success in applications, most of these methods are difficult to interpret and mathematical understanding of them is lacking. This paper studies the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. We establish asymptotic theories of maximum partial likelihood estimators and show that our nonparametric deep neural network estimator achieves the minimax optimal rate of convergence (up to a polylogarithmic factor). Moreover, we prove that the corresponding finite-dimensional estimator for treatment covariate effects is root-n consistent, asymptotically normal and attains semiparametric efficiency. Extensive simulation studies and analyses of real survival data sets show the proposed estimator produces confidence intervals with superior coverage as well as survival time predictions with superior concordance to actual survival times.

Statistical Inference for Counting Processes under Shape Heterogeneity

♦Yifei Sun[1] and Ying Sheng[2]
[1]Columbia University
[2]Chinese Academy of Sciences

Proportional rate models are among the most popular methods for analyzing the rate function of counting processes. Although providing a straightforward rate-ratio interpretation of covariate effects, the proportional rate assumption implies that covariates do not modify the shape of the rate function. When such an assumption does not hold, we propose describing the relationship between the rate function and covariates through two indices: the shape index and the size index. The shape index allows the covariates to flexibly affect the shape of the rate function, and the size index retains the interpretability of covariate effects on the magnitude of the rate function. To overcome the challenges in simultaneously estimating the two sets of parameters, we propose a conditional pseudolikelihood approach to eliminate the size parameters in shape estimation and an event count projection approach for size estimation. The proposed estimators are asymptotically normal with a root-n convergence rate. Simulation studies and an analysis of recurrent hospitalizations using SEER-Medicare data are conducted to illustrate the proposed methods.

Variable Selection for Interval-Censored Failure Time Data

Jianguo Sun

Interval-censored failure time data occur in many areas, including demographical studies, economic studies, medical studies and social sciences, and in different forms. This talk will discuss variable selection for such data and present some recently

developed tools.

Session 23INT2: Statistical Representative Points and Their Application in Statistical Inference

Remarks on Parameter Estimation of Mixture of Two Normal Distributions Based on Small Sample Sizes

♦Yinan Li and Kai-Tai Fang
BNU-HKBU United International College
TBC

Statistical Representative Points and their Application in Statistical Inference

♦Sirao Wang, Jiajuan Liang, Min Zhou and Huajun Ye
TBC

On the Non-Inferiority Mcnemar Test

Zhigang Zhang
Memorial Sloan-Kettering Cancer Center

The McNemar test is commonly used for comparing two dependent proportions arising from paired or matched individuals. In this study we thoroughly examine the McNemar test under the non-inferiority setting. We show that the usual conditional test bears a systematic bias while the unconditional test does not. We also derive sample size formula for practical usage when designing trials. Both theoretical considerations and numerical studies are carried out for establishing large sample properties and illustrating finite sample performance.

Session 23INT86: Statistical Methods for Robust Inference

Residual Projection for Quantile Regression in Vertically Partitioned Big Data

Ye Fan[1], Jr-Shin Li[2] and ♦Nan Lin[2]
[1]Capital University of Economics and Business
[2]Washington University in St. Louis

Traditional algorithms for quantile regression are no longer feasible for big data vertically stored in distributed environments, i.e. different variables are stored separately. While the popular alternating direction method of multipliers (ADMM) provides a viable computational solution, its slow convergence becomes a bottleneck when communication cost dominates local computational consumption, such as on Internet of Things (IoT) networks. Motivated by the residual projection technique, we propose an iterative parallel framework, PIQR, that converges faster and has a more secure data transmission plan, and establish its convergence property. Simulation studies show that both the ADMM-based method and the PIQR enjoy favorable estimation accuracy in distributed environments. While PIQR is inferior to the ADMM-based method at local computation, it requires much fewer iterations to achieve convergence, and hence significantly improves the overall computational efficiency when communication cost is the dominating factor. Moreover, PIQR transmits only data involving the residual information between different machines, and can better prevent the leakage of important data information compared with the ADMM-based method.

Tensor Response Quantile Regression with Neuroimaging Data

Bo Wei, ♦Limin Peng, Ying Guo, Amita Manatunga and Jennifer Stevens
Emory University

Collecting neuroimaging data in the form of tensors (i.e. multidimensional arrays) has become more common in mental health studies, driven by an increasing interest in studying the associations between neuroimaging phenotypes and clinical disease manifestation. Motivated by a neuroimaging study of post traumatic stress disorder (PTSD) from the Grady Trauma Project, we study a tensor response quantile regression framework, which enables novel analyses that confer a detailed view of the potentially heterogeneous association between a neuroimaging phenotype and relevant clinical predictors. We adopt a sensible low-rank structure to represent the association of interest, and propose a simple two-step estimation procedure which is easy to implement with existing software. We provide rigorous theoretical justifications for the intuitive two-step procedure. Simulation studies demonstrate good performance of the proposed method with realistic sample sizes in neuroimaging studies. We conduct the proposed tensor response quantile regression analysis of the motivating PTSD study to investigate the association between fMRI resting-state functional connectivity and PTSD symptom severity. Our results uncover non-homogeneous effects of PTSD symptoms on brain functional connectivity, which cannot be captured by existing tensor response methods.

Recursive Quantile Estimation: Non-Asymptotic Confidence Bounds

♦Likai Chen[1], Georg Keilbar[2] and Wei Biao Wu[3]
[1]Washington University in Saint Louis
[2]University of Vienna
[3]University of Chicago

This paper considers the recursive estimation of quantiles using the stochastic gradient descent (SGD) algorithm with Polyak-Ruppert averaging. The algorithm offers a compu- tationally and memory efficient alternative to the usual empirical estimator. Our focus is on studying the non-asymptotic behavior by providing exponentially decreasing tail prob- ability bounds under mild assumptions on the smoothness of the density functions. This novel non-asymptotic result is based on a bound of the moment generating function of the SGD estimate. We apply our result to the problem of best arm identification in a multi-armed stochastic bandit setting under quantile preferences.

An Empirical Bayes Method for Replicability Analysis of High-Dimensional Genomic Data

♦Yan Li[1], Xiang Zhou[2], Rui Chen[3], Xianyang Zhang[4] and Hongyuan Cao[5]
[1]Jilin University
[2]University of Michigan
[3]Baylor College of Medicine
[4]Texas A&M University
[5]Florida State University

Identifying replicable signals across multiple studies that examine the same features is a cornerstone of modern scientific research, which provides more reliable and stronger evidence for scientific findings. We focus on statistical methods for testing high-dimensional replicability null hypotheses across two studies regarding various genomic data. Due to the composite nature of replicability null hypotheses, most current methods either cannot control the false discovery rate or have low power. We develop an empirical Bayes method for identifying replicable signals from two high-dimensional genomic studies. We model the distribution of p-values across two studies with a four-group mixture model of the null and alternative distributions conditional on the joint hidden states. Heterogeneity is incorporated by assuming different alternative distributions for two studies. With an EM algorithm in combination with pool-adjacent-violator-algorithm, local false discovery rate regarding the replicability null across two studies can be estimated without any tuning parameter, producing different rankings of feature importance. By borrowing information across features and studies, the proposed method provides effective false discovery rate control and has a substantial power gain. We conduct extensive simulation studies and analyze data from spatially resolved transcriptomic studies to demonstrate the higher performance of proposed method.

Session 23INT76: Modern Approaches to Tackle Challenging Design Problems in Clinical Trials

Seamless Phase Ii/Iii Clinical Trials with Covariate Adaptive Randomization

Hongjian Zhu

There is an urgent need to evaluate new therapies in a time-sensitive and cost-effective manner. We propose the adaptive seamless phase II/III clinical trials with covariate adaptive randomization (CAR) to satisfy this need. CAR is one of the most popular designs in randomized controlled trials, enhancing covariance balance and ensuring valid treatment comparisons. However, it has several challenges: (1) the type I error rate of the commonly used Student's t-test following CAR can be inflated because of the seamless trials, but can also be decreased using CAR; (2) the complicated allocation mechanism induced by CAR causes extra difficulties to derive the asymptotic properties of a test procedure; and (3) previous theoretical studies of seamless trials rely mainly on the assumption of complete randomization, a procedure rarely used in real trials. We establish a theoretical foundation for adaptive seamless phase II/III trials with CAR. We also propose an approach that is easy to implement in order to control the type I error rate and improve the power when using Student's t-test. This important step will promote the application of this procedure.

Optimizing the Design of Pediatric Pharmacokinetic Trials – a Case Study Evaluating a Novel Drug for Treatment of Multidrug-Resistant Tuberculosis (Mdr-Tb) in Children with and without Hiv

♦Grace Montepiedra[1], Elin Svensson[2], Weng Kee Wong[3] and Andrew Hooker[2]
[1]Harvard T.H. Chan School of Public Health
[2]Uppsala University
[3]UCLA

Pharmacokinetic (PK) studies in children are usually small and have ethical constraints due to the medical complexities of drawing blood in this special population. Tapping on the availability of population PK models, extrapolated from models based on adult data to children, there is increasing interest in the use of optimal design methodology to design PK sampling schemes that maximize information using a small sample size and limited

number of sampling times per dosing period. In this investigation we use the novel TB drug Delamanid, which has a very complex model structure, as a case study to show the strengths and challenges of application of optimal design methodology in this context, eventually leading to highly efficient designs for estimation of PK parameters with a limited number of sampling measurements. Using developed population PK models based on available data from adults with and without HIV, and limited data on children without HIV, competing designs were derived and assessed based on robustness to model uncertainty when extrapolated to children living with HIV. We show how this is implemented using the R package PopED developed by one of our collaborators.

Metaheuristics for Designing Efficient Biomedical Studies

Weng Kee Wong

University of California at Los Angeles

This talk presents an overview of nature-inspired metaheuristic algorithms and their emerging use in statistical research with a focus on biomedical applications. These algorithms are virtually assumptions-free, flexible and are general optimization tools. To fix ideas, I discuss an exemplary such algorithm called particle swarm optimization, which is one of the most popular and widely used nature-inspired algorithms. It has many variants and I discuss its applications to solve different types of statistical problems. I will present various applications to tackle biomedical design problems, and if time permits, I will provide live demonstrations.

Enhancing the Performance of Metaheuristic Algorithms by Appropriate Noise Addition

♦*Kwok Pui Choi*[1], *Enzio Hai Hong Kam*[1], *Xin Tong*[1] *and Weng Kee Wong*[2]

[1]National University of Singapoe

[2]UCLA

Nature-inspired swarm-based algorithms are increasingly applied to tackle high-dimensional and complex optimization problems across disciplines. They are general purpose optimization algorithms, easy to implement and almost assumption-free. Some common drawbacks of these algorithms are their premature convergence and the solution found may not be a global optimum. We propose a general, yet simple and effective strategy, called heterogeneous Perturbation-Projection (HPP), to enhance an algorithm's exploration capability so that our sufficient convergence conditions are guaranteed to hold and the algorithm converges almost surely to a global optimum. HPP applies stochastic perturbation on half of the swarm agents and then project all agents onto the set of feasible solutions. In this talk, we apply this approach to a widely used nature-inspired swarm-based optimization algorithm: particle swarm optimization (PSO). We report extensive numerical experiments that demonstrate the HPP-modified PSO outperforms the original PSO version. Applications of HPP to other optimization algorithms: Bat algorithm, Ant Colony Optimization and Competitive Swarm Optimization, will be briefly sketched. This talk is based on a joint work with Enzio Kam, Xin Tong and Weng Kee Wong.

Session 23INT57: Recent Developments in Optimal Treatments and High Dimensional Data Analysis

A Quasi-Optimal Dose-Finding Approach in Infinite Horizon Dynamic Treatment Regime

Yuhan Li[1], *Wenzhuo Zhou*[2] *and* ♦*Ruoqing Zhu*[1]
[1]University of Illinois Urbana Champaign
[2]University of California Irvine

Many real-world reinforcement learning (RL) applications require making decisions in continuous action environments. In particular, determining the optimal dose level is vital in developing dynamic treatment regimes. One challenge in adapting existing RL algorithms to medical applications, however, is that the popular infinite support stochastic policies, e.g., Gaussian policy, may assign riskily high dosages and harm patients seriously. Hence, it is important to induce a policy class whose support only contains near-optimal actions and shrink the action-searching area for effectiveness and reliability. To achieve this, we develop a novel *quasi-optimal learning algorithm*, which can be easily optimized in off-policy settings with guaranteed convergence under general function approximations. Theoretically, we analyze the consistency, sample complexity, adaptability, and convergence of the proposed algorithm. We evaluate our algorithm with comprehensive simulated experiments and a dose suggestion real application to Ohio Type 1 diabetes dataset.

Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

♦*Yunan Wu*[1], *Lan Wang*[2] *and Haoda Fu*[3]
[1]University of Texas at Dallas
[2]University of Miami
[3]Eli Lilly and Company

This article develops new tools to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modeling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimension and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound. We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference for the effect of a group of variables on optimal decision making. The advantage of honest inference is that it does not require the initial estimator to achieve perfect model selection and does not require the zero and nonzero effects to be well-separated. We also propose an efficient algorithm for estimation. Our simulations suggest satisfactory performance. An example from a diabetes study illustrates the real application.

A Unified Generalization of Inverse Regression via Adaptive Column Selection

Yin Jin and ♦*Wei Luo*
Zhejiang University
Higher-order inverse regression methods are commonly known as more powerful sufficient dimension reduction (SDR) meth-

ods than the popularly used sliced inverse regression (SIR) in the population level. However, due to the convention of essentially conducting singular value decomposition on the ambient candidate matrices, these methods suffer from the excessive number of parameters in the sample level and have not been systematically generalized under the high-dimensional settings like SIR. In this paper, we break the convention of using the ambient candidate matrices in these methods, and instead apply a novel column-selection strategy on their candidate matrices that substantially lowers down the working number of parameters to being comparable with SIR. Then, for the first time of the literature, we generalize the higher-order inverse regression methods, as well as their ensembles, towards sparsity under the high-dimensional settings in a uniform manner. The dimension of the predictor is allowed to diverge with the sample size in nearly an exponential order, and no additional restrictions are imposed on the data other than those commonly seen in the high-dimensional literature. For completeness of theory, we also study the column-selection strategy towards the estimation efficiency under the conventional low-dimensional settings. These results are illustrated by simulation studies and a real data application at the end.

Regional Quantile Regression for Multiple Responses

Eun Ryung Lee
SungKyunKwan University

In this article, we study high-dimensional multiple response quantile regression model for an interval of quantile levels, in which a common set of covariates is used to analyze multiple responses simultaneously. We assume that the underlying quantile coefficient matrix is simultaneously element-wise and row-wise sparse. We address high dimensional issues to identify globally relevant variables for multiple responses when any th conditional quantile is considered, where Δ, and Δ is an interval of quantile levels of interest. We develop a novel penalized globally concerned quantile regression with double group Lasso penalties and propose an information criterion for penalty parameter choice. We prove that the proposed method consistently selects both element-wise and row-wise sparsity patterns of the regression coefficient matrix function and that it achieves the oracle convergence rate. Numerical examples and applications to Cancer Cell Line Encyclopedia data illustrate the advantages of the proposed method over separate penalized quantile regression on each response.

Session 23INT69: Recent Developments in Experimental Designs

Optimal Designs in Mixed Models for Repeated Measurements

♦*Xiaojian Xu[1] and Sanjoy Sinha[2]*
[1]Brock University
[2]Carleton University

We discuss the construction of optimal designs for linear mixed models with covariates when involving repeated measurements. Random effects are employed to accommodate the clusters. We consider both the treatment effects as well as continuous covariates in the model. The goal of the designs is to optimally select the levels of covariates as well as the proportions of the sample units allocated to each treatment within a given total sample size. Both D- and A-optimality are chosen to be the design criteria. Although the estimators can be given with analytic forms if normality is assumed, the optimal designs depend on the unknown parameters involved in the variance components. Therefore, we apply both two-stage and sequential approaches.

A Machine Learning Perspective for Optimal Design using Tight Mutual Information

♦*Xinwei Deng and Qing Guo*
Department of Statistics, Virginia Tech

Effective data collection is very important in data science. In this talk, we consider a machine learning perspective for optimal experimental design. The proposed work considers the Bayesian optimal experimental design (BOED) framework based on maximizing the mutual information (MI) between data and model parameters. However, directly applying existing MI estimators may not work since they rely on explicit knowledge of the model likelihood. To overcome these limitations, we revisit the popular variational MI bounds from the lens of statistical modeling and convex optimization. Our investigation leads to a novel, simple, and powerful contrastive MI estimator for optimal experimental design. The performance of the proposed method is evaluated by both the likelihood-free experimental design and the amortized sequential experimental design.

A Bayesian Approach to Process Optimization on Data with Multi-Stratum Structure

Xiaohua Liu[1], ♦Po Yang[1] and Chang-Yun Lin[2]
[1]University of Manitoba, Canada
[2]National Chung Hsing University, Taiwan

Multistratum design arises naturally in industrial experiments due to the inconvenient and impractical completely randomization. Most research has concentrated on finding optimal multistratum designs that have high efficiencies in parameter estimation. Accounting for the model uncertainty, we apply the Bayesian model averaging method and predictive approach to investigate the optimization problem for data with multistratum structure. With the posterior probabilities of models as weights, we consider the weighted average of the predictive densities of the response over all potential models. The goal of the optimization is to identify the values of the factors that result in a maximum probability of a response between a given range. The method is illustrated with two examples.

Generalized Bayesian d-Optimal Supersaturated Multistratum Designs

Chang-Yun Lin
NCHU

Supersaturated designs are useful in the initial stage of experiments to identify important factors from many of interest with a small number of runs. Traditional supersaturated designs were mainly constructed for completely randomized experiments, which have single-stratum structures. They cannot be used for experiments that have multistratum structures, such as the split-plot, strip-plot, and staggered-level experiments. How to construct supersaturated multistratum designs for complex experiments has gained much attention recently. In this paper, we consider the situation in which the experimenters have prior knowledge of which factors are more likely to be important (called the primary factors) than the others (called the potential factors). By taking primary and potential factors into account, we propose an approach using the generalized Bayesian D (GBD) criterion to con- struct a new class of supersaturated

multistratum designs. The GBD-optimal supersaturated multistratum designs provide guidelines on how to assign factors to the designs, which enhances efficiency on identifying active factors. A case study shows that the proposed supersaturated design (32 runs with 19 factors) is as effective as the full 2^6 factorial design (64 runs with 6 factors) to identify important factors in a battery cell experiment.

Session 23INT7: Methodology Advances in Analyzing High-Throughput Genomic and Genetics Data

Summit-Fa: a New Resource for Improved Transcriptome Imputation using Functional Annotations

Melton Melton[1], Zichen Zhang[1] and ♦Chong Wu[2]
[1]Florida State University
[2]UT MD Anderson Cancer Center

Transcriptome-wide association studies (TWAS) integrate gene expression prediction models and genome-wide association studies (GWAS) to identify gene–trait associations. The power of TWAS is determined by the sample size of GWAS and the accuracy of the expression prediction model. Here, we present a new method, the Summary-level Unified Method for Modeling Integrated Transcriptome using Functional Annotations (SUMMIT-FA), that improves the accuracy of gene expression prediction by leveraging functional annotation resources and a large expression quantitative trait loci (eQTL) summary-level dataset. We build gene expression prediction models using SUMMIT-FA with a comprehensive functional database MACIE and the eQTL summary-level data from the eQTLGen consortium. By applying the resulting models to GWASs for 24 complex traits and exploring it through a simulation study, we show that SUMMIT-FA improves the accuracy of gene expression prediction models in whole blood, identifies significantly more gene-trait associations, and improves predictive power for identifying "silver standard" genes compared to several benchmark methods.

Individual-Specific Reference Panel Recovery Improves Cell-Type-Specific Inference

♦*Hao Feng[1], Guanqun Meng[1] and Qian Li[2]*
[1]Case Western Reserve University
[2]St. Jude Children's Research Hospital

We propose a statistical algorithm ISLET to infer individual-specific and cell-type-specific transcriptome reference panels. ISLET properly models the repeatedly measured bulk gene expression data, to optimize the usage of shared information within the same subject. ISLET is the first available method to achieve individual-specific reference estimation. Using simulation studies, we show favorable performance of ISLET in the reference estimation and downstream cell-type-specific differentially expressed genes testing. We apply ISLET on longitudinally profiled transcriptomes in blood samples from a large observational study of young children, and confirmed the cell type-related gene signatures for pancreatic islet autoantibody. ISLET is available at https://bioconductor.org/packages/ISLET.

A Cofunctional Grouping-Based Approach for Non-Redundant Feature Gene Selection in Unannotated Single-Cell Rna-Seq Analysis

Xiaobo Sun
Zhongnan University of Laws and Economics

TBA

Session 23INT85: New Regression, Prediction, and Screening in Clinical Trials

Covidscreen: a Web App and r Package for Assessing Asymptomatic Covid 19 Testing Strategies

♦*Li Tang, Jesse Smith, Yiwang Zhou, Motomi Mori and Akshay Sharma*
St. Jude Children's Research Hospital

TBC

Regression Analysis for Covariate-Adaptive Randomization: a Robust and Efficient Inference Perspective

♦*Wei Ma[1], Fuyi Tu[1] and Hanzhong Liu[2]*
[1]Renmin University of China
[2]Tsinghua University

Linear regression is arguably the most fundamental statistical model; however, the validity of its use in randomized clinical trials, despite being common practice, has never been crystal clear, particularly when stratified or covariate-adaptive randomization is used. In this article, we investigate several of the most intuitive and commonly used regression models for estimating and inferring the treatment effect in randomized clinical trials. By allowing the regression model to be arbitrarily misspecified, we demonstrate that all these regression-based estimators robustly estimate the treatment effect, albeit with possibly different efficiency. We also propose consistent non-parametric variance estimators and compare their performances to those of the model-based variance estimators that are readily available in standard statistical software. Based on the results and taking into account both theoretical efficiency and practical feasibility, we make recommendations for the effective use of regression under various scenarios. For equal allocation, it suffices to use the regression adjustment for the stratum covariates and additional baseline covariates, if available, with the usual ordinary-least-squares variance estimator. For unequal allocation, regression with treatment-by-covariate interactions should be used, together with our proposed variance estimators. These recommendations apply to simple and stratified randomization, and minimization, among others. We hope this work helps to clarify and promote the usage of regression in randomized clinical trials.

Genomics, Machine Learning, and Fracture Risk Assessment

Qing Wu

TBC

Low-Rank Latent Matrix-Factor Prediction Modeling for Generalized High-Dimensional Matrix-Variate Regression

Catherine Liu
The Hong Kong Polytechnic University

Our talk is motivated by diagnosing the COVID-19 disease using 2D image biomarkers from computed tomography (CT) scans. We propose a novel latent matrix-factor regression model to predict responses that may come from an exponential distribution family, where covariates include high-dimensional matrix-variate biomarkers. A latent gen- eralized matrix regression (LaGMaR) is formulated, where the latent predictor is a low-dimensional matrix factor score extracted from the low-rank signal of the matrix variate through a cutting-edge matrix factor model. Unlike the general spirit of penal- izing vectorization plus the necessity of tuning parameters in the literature, instead, our prediction

modeling in LaGMaR conducts dimension reduction that respects the geometric characteristic of intrinsic two-dimensional structure of the matrix co- variate and thus avoids iteration. This greatly relieves the computation burden, and meanwhile maintains structural information so that the latent matrix factor feature can perfectly replace the intractable matrix-variate owing to high-dimensionality. The estimation procedure of LaG-MaR is subtly derived by transforming the bilin- ear form matrix factor model onto a high-dimensional vector factor model, so that the method of principle components can be applied. We establish bilinear-form con- sistency of the estimated matrix coefficient of the latent predictor and consistency of prediction. The proposed approach can be implemented conveniently. Through simulation experiments, the prediction capability of LaGMaR is shown to outper- form some existing penalized methods under diverse scenarios of generalized matrix regressions. Through the application to a real COVID-19 dataset, the proposed approach is shown to predict efficiently the COVID-19

Session 23INT10: Advances in Survival Analysis and Its Applications

Optimal Cut-Point of Marker for Early Disease Detection
♦*Cuiling Wang, Mindy Katz, Carol Derby and Richard Lipton*
Albert Einstein College of Medicine

Selection of a cut-point for a diagnostic marker of disease to identify those at high risk of developing the disease in the future is critically important in clinical practice and research. To screen currently healthy individuals at risk for developing disease within a certain follow-up time period, an optimal threshold based on rules such as Youden's index can be obtained through the time-dependent receiver operating characteristic (ROC) analysis. However, little is known regarding the property of the optimal cut-point. We investigate the property of the time-dependent optimal cut-point based on Youden's index and provide methods to directly estimate the optimal cuts based on the property using commonly used survival models. The method is applied to screening for pre-clinical Alzheimer's dementia using a well-established memory test in the Einstein Aging Study. Simulation studies are performed to evaluate the direct estimates as well as those obtained from the time-dependent ROC analysis. We show that when a marker is associated with disease such that a higher value indicates higher risk of disease incidence, the optimal cut-point decreases over time. The estimates perform well under correctly specified survival models, are similar to estimates obtained from the time-dependent ROC analysis using the same model-based estimators, and are more efficient than estimates obtained from the time-dependent ROC analysis using non-parametric or semi-parametric methods.

Bayesian and Influence Function Based Empirical Likelihoods for Inference of Sensitivity to the Early Diseased Stage in Diagnostic Tests.
♦*Gengsheng Qin, Shuangfei Shi and Yan Hai*
Georgia State University

In practice, a disease process might involve three ordinal diagnostic stages: the normal healthy stage, the early stage of the disease, and the stage of full development of the disease. Early detection is critical for some diseases since it often means an optimal time window for therapeutic treatments of the diseases. In this study, we propose a new influence function-based empirical

likelihood method and Bayesian empirical likelihood methods to construct confidence/credible intervals for the sensitivity of a test to patients in the early diseased stage given a specificity and a sensitivity of the test to patients in the fully diseased stage. Numerical studies are performed to compare the finite sample performances of the proposed approaches with existing methods. The proposed methods are shown to outperform existing methods in terms of coverage probability. A real dataset from the Alzheimer's Disease Neuroimaging Initiative (ANDI) is used to illustrate the proposed methods.

using Auxiliary Information for Estimation with Left Truncated Data
Yidan Shi[1], ♦Leilei Zeng[2], Mary E. Thompson[2] and Suzanne Tyas[2]
[1]University of Pennsylvania
[2]University of Waterloo

In life history studies one often encounters situations where individuals in a population are eligible to enroll only if the response time does not exceed an associated censoring time, which leads to the so called left truncated lifetime data. While auxiliary information for the truncated individuals from the same or similar cohorts may be available, challenges arise due to the practical issue of accessibility of individual-level data and taking account of various sampling conditions for different cohorts. We propose a likelihood-based method for incorporating auxiliary data to eliminate the bias due to left-truncation and improve efficiency. Simulation results and an application to data from a longitudinal study of aging are given.

Mean Residual Life Cure Models for Right-Censored Data Subject to Length-Biased Sampling
Chyong-Mei Chen[1], Hsin-Jen Chen[1] and ♦Yingwei Peng[2]
[1]National Yang Ming Chiao Tung University, Taiwan
[2]Queen's University, Canada

We propose a semiparametric mean residual life mixture cure model for right-censored survival data with a cured fraction. The model employs the proportional mean residual life model to describe the effects of covariates on the mean residual time of uncured subjects and the logistic regression model to describe the effects of covariates on the cure rate. We develop estimating equations to estimate the proposed cure model for the right-censored data with and without length-biased sampling, the latter is often found in prevalent cohort studies. In particular, we propose two estimating equations to estimate the effects of covariates in the cure rate and a method to combine them to improve the estimation efficiency. The consistency and asymptotic normality of the proposed estimates are established. The finite sample performance of the estimates is confirmed with simulations. The proposed estimation methods are applied to a clinical trial study on melanoma and a prevalent cohort study on early-onset type 2 diabetes mellitus.

Session 23INT26: Statistical Modeling of Complex-Data with Censoring and Measurement Errors

Novel Empirical Likelihood Inference for the Mean Difference with Right-Censored Data
Kangni Alemdjrodo[1] and ♦Yichuan Zhao[2]
[1]Purdue University
[2]Georgia State University

This paper focuses on comparing two means and finding a confidence interval for the difference of two means with right-censored data using the empirical likelihood method combined with the i.i.d. random functions representation. Some early researchers proposed empirical likelihood-based confidence intervals for the mean difference based on right-censored data using the synthetic data approach. However, their empirical log-likelihood ratio statistic has a scaled chi-squared distribution. To avoid the estimation of the scale parameter in constructing confidence intervals, we propose an empirical likelihood method based on the i.i.d. representation of Kaplan–Meier weights involved in the empirical likelihood ratio. We obtain the standard chi-squared distribution. We also apply the adjusted empirical likelihood to improve coverage accuracy for small samples. We investigate a new empirical likelihood method, the mean empirical likelihood, within the framework of our study. Via extensive simulations, the proposed empirical likelihood confidence interval has better coverage accuracy than those from existing methods. Finally, our findings are illustrated with a real data set.

Semiparametric Regression Analysis of Partly Interval-Censored Failure Time Data with Application to an Aids Clinical Trial

◆Qingning Zhou[1], Yanqing Sun[1] and Peter Gilbert[2]
[1]University of North Carolina at Charlotte
[2]Fred Hutchinson Cancer Center

Failure time data subject to various types of censoring commonly arise in epidemiological and biomedical studies. Motivated by an AIDS clinical trial, we consider regression analysis of failure time data that include exact and left-, interval-, and/or right-censored observations, which are often referred to as partly interval-censored failure time data. We study the effects of potentially time-dependent covariates on partly interval-censored failure time via a class of semiparametric transformation models that includes the widely used proportional hazards model and the proportional odds model as special cases. We propose an EM algorithm for the nonparametric maximum likelihood estimation and show that it unifies some existing approaches developed for traditional right-censored data or purely interval-censored data. In particular, the proposed method reduces to the partial likelihood approach in the case of right-censored data under the proportional hazards model. We establish that the resulting estimator is consistent and asymptotically normal. In addition, we investigate the proposed method via simulation studies and apply it to the motivating AIDS clinical trial.

Analysis of Noisy Survival Data with Graphical Proportional Hazards Measurement Error Model

◆Grace Yi[1] and Li-Pang Chen[2]
[1]University of Western Ontario
[2]National Chengchi University

In survival data analysis, the Cox proportional hazards (PH) model is perhaps the most widely used model to feature the dependence of survival times on covariates. While many inference methods have been developed under such a model or its variants, those models are not adequate for handling data with complex structured covariates. High-dimensional survival data often entail several features: (1) many covariates are inactive in explaining the survival information, (2) active covariates are associated in a network structure, and (3) some covariates are error-contaminated. To handle such kinds of survival data, we propose graphical PH measurement error models and develop inferential procedures for the parameters of interest. Our proposed models significantly enlarge the scope of the usual Cox PH model and have great flexibility in characterizing survival data. Theoretical results are established to justify the proposed methods. Numerical studies are conducted to assess the performance of the proposed methods.

Session 23INT14: Advances in Statistical Methods for Machine Learning

Mdp2 Forest: a Constrained Continuous Multi-Dimensional Policy Optimization Approach for Short-Video Recommendation

Fan Zhou

In the ecology of short video platforms, the optimal exposure proportion of each video category is crucial to guide recommendation systems and content production in a macroscopic way. Though extensive studies on recommendation systems are devoted to providing the most well-matched videos for each view request, fitting the data without considering inherent biases such as selection bias and exposure bias will result in serious issues. In this paper, we formalize the exposure proportion strategy as a policy-making problem with multi-dimensional continuous treatment under certain constraints from a causal inference point of view. We propose a novel ensemble policy learning method based on causal trees, called Maximum Difference of Preference Point Forest (MDP2 Forest), which overcomes the shortcomings of existing policy learning approaches. Experimental results on both simulated and synthetic datasets show the superiority of our algorithm compared to other policy learning or causal inference methods in terms of the treatment estimation accuracy and the mean regret. Furthermore, the proposed MDP2 Forest method can also adapt to a wide range of business settings such as imposing different kinds of constraints on the multi-dimensional treatment.

Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling

Qi Xu[1], Yubai Yuan[2], Junhui Wang[3] and ◆Annie Qu[1]
[1]UC Irvine
[2]Penn State
[3]CUHK

Crowdsourcing has emerged as an alternative solution for collecting large scale labels. However, the majority of recruited workers are not domain experts, so their contributed labels could be noisy. In this paper, we propose a two-stage model to predict the true labels for multicategory classification tasks in crowdsourcing. In the first stage, we fit the observed labels with a latent factor model and incorporate subgroup structures for both tasks and workers through a multi-centroid grouping penalty. Group-specific rotations are introduced to align workers with different task categories to solve multicategory crowdsourcing tasks. In the second stage, we propose a concordance-based approach to identify high-quality worker subgroups who are relied upon to assign labels to tasks. In theory, we show the estimation consistency of the latent factors and the prediction consistency of the proposed method. The simulation studies show that the proposed method outperforms the existing competitive methods, assuming the subgroup structures within tasks and

workers. We also demonstrate the application of the proposed method to real world problems and show its superiority.

Online Statistical Inference for Matrix Contextual Bandit

Qiyu Han, ♦*Will Wei Sun and Yichen Zhang*
Purdue University

Contextual bandit has been widely used for sequential decision-making based on the current contextual information and historical feedback data. In modern applications, such context format can be rich and can often be formulated as a matrix. Moreover, while existing bandit algorithms mainly focused on reward-maximization, less attention has been paid to the statistical inference. To fill in these gaps, in this work we consider a matrix contextual bandit framework where the true model parameter is a low-rank matrix, and propose a fully online procedure to simultaneously make sequential decision-making and conduct statistical inference. The low-rank structure of the model parameter and the adaptivity nature of the data collection process makes this difficult: standard low-rank estimators are not fully online and are biased, while existing inference approaches in bandit algorithms fail to account for the low-rankness and are also biased. To address these, we introduce a new online doubly-debiasing inference procedure to simultaneously handle both sources of bias. In theory, we establish the asymptotic normality of the proposed online doubly-debiased estimator and prove the validity of the constructed confidence interval. Our inference results are built upon a newly developed low-rank stochastic gradient descent estimator and its non-asymptotic convergence result, which is also of independent interest.

Session 23INT11: Recent Developments in Survival Analysis with High-Dimensional or Longitudinal Covariates

Fast Lasso-Type Safe Screening for Fine-Gray Competing Risks Model with Ultrahigh Dimensional Covariates

♦*Hong Wang*[1] *and Gang Li*[2]
[1]Central South University
[2]UCLA

In recent years, there has been growing interest in developing safe feature elimination (SAFE) algorithms for high dimensional and/or large scale Lasso-type problems. However, its application to survival analysis is rather limited. In this paper, we develop a SAFE algorithm for the popular Fine-Gray proportional sub-distribution hazards (PSH) model for competing risks survival data, in which subjects may experience from multiple mutually exclusive failures subject to right censoring. Specifically, we derive SAFE rules for the Lasso and adaptive Lasso Fine-Gray competing risks model using convex optimization theory. We evaluate the performance of the proposed procedure in terms of screening efficiency and safety, runtime, and prediction accuracy on multiple simulation datasets and a public bladder cancer dataset. The empirical results show that the proposed approach is effective with respect to all the above criteria, especially in terms of computational efficiency.

Sure Joint Screening for High Dimensional Cox's Proportional Hazards Model under the Case-Cohort Design

♦*Yi Liu*[1] *and Gang Li*[2]
[1]Ocean University of China
[2]University of California at Los Angeles

This paper develops a sure joint feature screening method for the case-cohort design with ultrahigh dimensional covariates. Our method is based on a sparsity-restricted Cox's proportional hazards model. An iterative reweighted hard thresholding algorithm is proposed to approximate the sparsity-restricted pseudo-partial likelihood estimator for joint screening. We show rigorously that our method possesses the sure screening property, with the probability of retaining all relevant covariates tending to 1 as the sample size goes to infinity. Our simulation results demonstrate that the proposed procedure has substantially improved screening performance over some existing feature screening methods for the case-cohort design especially when some covariates are jointly correlated, but marginally uncorrelated, to the event time outcome. A real data illustration is provided on a breast cancer data with high dimensional genomic covariates. We have implemented the proposed method using Matlab and made it available to readers through Github.

Kernel Meets Sieve: Transformed Hazards Models with Sparse Longitudinal Covariates

Hongyuan Cao[1]*, Dayu Sun*[2]*, Zhuowei Sun*[3] *and* ♦*Xingqiu Zhao*[4]
[1]Florida State University
[2]Emory University
[3]Jilin University
[4]Hong Kong Polytechnic University

In survival studies, time-dependent covariates usually cannot be monitored continuously but are measured only at several discrete time points, generating sparse longitudinal covariates. The covariate observation sparsity precludes classic hazard-based survival analysis due to missing necessary covariate information. Existing methods to cope with sparse longitudinal covariates either require restrictive assumptions and high computational resources or lack rigorous theoretical justification. We propose to combine kernel-weighted log-likelihood and sieve maximum log-likelihood estimation under a flexible class of transformed hazards models. The proposed method enjoys simple computation and needs minimal assumptions. We establish the asymptotic properties of the proposed estimators whose proofs contribute to a rigorous theoretical framework for general kernel-weighted semiparametric M-estimators. Numerical studies confirm our theoretical results and show that the proposed method outperforms existing methods. An application to a recent COVID-19 study in Wuhan illustrates the practical utility of our proposal.

Session 23INT103: Invited Session on Lifetime Data Analysis

Kullback-Leibler-Based Relative Risk Models for Integration of Published Survival Models with New Dataset

♦*Kevin He and Di Wang*
University of Michigan

Prediction of time-to-event data often suffers from rare event rates, small sample sizes, high dimensionality and low signal-to-noise ratios. Incorporating published prediction models from large-scale studies is expected to improve the performance of prognosis prediction on internal small-sized time-to-event data. To account for challenges including heterogeneity, data sharing, and privacy constraints, I will introduce a data integration procedure based on Kullback-Leibler divergence, which measures the discrepancy between the published models and the internal

dataset. The proposed procedure is computationally efficient for high-dimensional problems and can be easily implemented with various machine learning methods. Asymptotic properties and simulation results show the advantage of the proposed method compared with those solely based on the internal data. We apply the proposed method to improve prediction performance on a kidney transplant dataset from a local hospital by integrating this small-scale dataset with published survival models obtained from the national transplant registry.

Profile Optimum Planning for Degradation Analysis

♦Chien-Yu Peng and Ya-Shan Cheng
Academia Sinica

Degradation tests are widely used to assess the lifetime information of highly reliable products. Before conducting a degradation test, the fundamental issues in regard to decision variables are to determine how many sample sizes are needed, how long samples need to be tested, and how many measurements are taken, particularly in early experiments with limited budgets. By minimizing the approximate variance of the estimated qth quantile of the product's lifetime distribution with a cost constraint, profile optimum planning (POP) is proposed to provide a systematic solution to these decision variables under mild conditions. Based on the derived theoretical results, the use of flow charts efficiently simplifies complex optimization problems in practical applications. In addition, sensitivity analysis is studied to elucidate the effect of how the uncertainty in POP can be divided and allocated to the experimental costs and model parameters.

Regression Analysis of Serial Gap Time Data with Recurrent and Terminal Events via Additive Hazards Models

Yong-Chen Huang and ♦Shu-Hui Chang
National Taiwan University

The long-term course of a chronic disease usually involves recurrent events and a terminal event. In epidemiological and medical studies, serial gap times between consecutive events are the natural outcomes of interest in which recurrent and terminal events are the primary endpoints. We introduce a series of cause-episode-specific hazard functions for the serial gap times by incorporating linear covariate effects, which may depend on the cause and episode of event. Semiparametric estimation methods for cause-episode-specific covariate effects are developed under various situations of censoring by allowing arbitrary pattern of association between serial gap times. Various tests are further constructed for testing the equivalence of covariate effects over different causes and episodes. Simulation studies and one real data set are presented to illustrate the proposed methods.

Surrogate Marker Assessment of Covid-19 Vaccine Efficacy using Mediation Analyses in a Case-Cohort Design

♦Yen-Tsung Huang, Jih-Chang Yu and Jui-Hsiang Lin
Academia Sinica

The identification of surrogate markers for gold standard outcomes in clinical trials enables future cost-effective trials that target the identified markers. Motivated by a COVID-19 vaccine trial, we propose methods of assessing the surrogate markers for a time-to-event outcome in a case-cohort design by using mediation analyses. We decomposed the vaccine effect on COVID-19 risk into an indirect effect (the effect mediated through the surrogate marker such as neutralizing antibodies), and a direct effect (the effect not mediated by the marker), and we pro-

pose that the mediation proportions are surrogacy indices. We employed weighted estimating equations derived from nonparametric maximum likelihood estimators (NPMLEs) under semiparametric probit models for the time-to-disease outcome. We plugged in the weighted NPMLEs to construct estimators for the aforementioned causal effects and surrogacy indices, and we determined the asymptotic properties of the proposed estimators. Finite sample performance was evaluated in numerical simulations. Applying the proposed mediation analyses to a mock COVID-19 vaccine trial data, we found that 84.2% of the vaccine efficacy was mediated by 50% pseudovirus neutralizing antibody.

Session 23INT34: Statistical Inference with Nonparametric and Semiparametric Methods

Paired or Partially Paired Two-Sample Tests with Unordered Samples

Yudong Wang[1], ♦Yanlin Tang[2] and Zhisheng Ye[1]
[1]National University of Singapore
[2]East China Normal University

In paired two-sample tests for mean equality, it is common to encounter unordered samples in which subject identities are not observed or unobservable, and it is impossible to link the measurements before and after treatment. The absence of subject identities masks the correspondence between the two samples, rendering existing methods inapplicable. In this paper, we propose two novel testing approaches. The first splits one of the two unordered samples into blocks and approximates the population mean using the average of the other sample. The second method is a variant of the first, in which subsampling is used to construct an incomplete U-statistic. Both methods are affine invariant and can readily be extended to partially paired two-sample tests with unordered samples. Asymptotic null distributions of the proposed test statistics are derived and the local powers of the tests are studied. Comprehensive simulations show that the proposed testing methods are able to maintain the correct size, and their powers are comparable to those of the oracle tests with perfect pair information. Four real examples are used to illustrate the proposed methods, in which we demonstrate that naive methods can yield misleading conclusions.

Estimation and Inference for Ultra-High Dimensional Quasi-Likelihood Models Based on Data Splitting

Xuejun Jiang
Southern University of Science and Technology

In this article, we develop a valid weighted estimation and inference framework for ultra-high dimensional quasi-likelihood models. The weighted estimator is obtained by minimizing the variance function. We split the full data into two subset, conduct the model selection on one subset and compute the maximum quasi-likelihood estimator on the other subset. Then we aggregate the two estimator with the optimal weighted matrixes to form the final weighted estimator. With the weighted estimator, we construct the confidence intervals for the group components of the regression vector, and the Wald test for the linear structure of the group components. Theoretically, we establish the asymptotic normality of the weighted estimator, and the asymptotic χ^2-distribution of the corresponding Wald test without assuming model selection consistency. Advantages of

the proposed tests are highlighted via theoretical and empirical comparison to some competitive tests, which guarantees that our proposed estimation and inference framework is locally optimal. In addition, when the selection consistency is achieved, we prove that the proposed Wald test is asymptotically identically distributed as the oracle tests in the sense that it knows the support of regression vector. Extensive simulations demonstrate more favorable finite sample performance of the proposed tests. An application to Arcene cancer data illustrates the use of our proposed methodology.

Dimension Reduction for Functional Time Series Model

♦*Guochang Wang and Zenyao Wen*

Functional time series model is the most widely studied in the recently years and functional data is infinite dimensional, then dimension reduction is crucial for functional time series. However, most of the existing dimension reduction methods (such as the functional principal component and fixed basis expansion) is unsupervised and these dimension reduction approaches usually lead to information loss. Then, a supervised dimension reduction method is urgently need for functional time series model. Functional sufficient dimension reduction method is a supervised approach and can sufficiently exploit the regression structure information, which leads to little information loss. The most popular functional sufficient dimension reduction methods is the functional sliced inverse regression (FSIR), but it can not be applied in functional time series model directly. In the presented paper, we consider a functional time series model where the response is a scalar time series and the explaining variable is functional time series. We combine the FSIR and blind source separation methods to propose a novel supervised dimension reduction method for this regression model. Furthermore, we propose a novel selected strategies to select the dimensionality of dimension reduction space and the lags of the functional time series. Lastly, Numerical studies including simulation studies and one real data analysis are demonstrated the usefulness of the proposed methods.

Regression Analysis of Partially Linear Transformed Mean Residual Life Models

♦*Haijin He*[1], *Jingheng Cai*[2] *and Xinyuan Song*[3]

[1]Shenzhen University
[2]Sun-Yat sen University
[3]The Chinese University of Hong Kong

We propose a novel class of partially linear transformed mean residual life (TMRL) models to investigate linear and nonlinear covariate effects on survival outcomes of interest. A martingale-based estimating equation approach with global and kernel-weighted local estimating equations is developed to estimate the parametric and nonparametric components. Unlike the existing inverse probability of censoring weighting estimating equation approach on TMRL models, the newly proposed method avoids estimating or modeling the distribution of the censoring time, thereby enhancing model capability and computational efficiency. Furthermore, we establish the asymptotic properties for the estimators of parametric and nonparametric components and develop an efficient iterative algorithm to implement the proposed procedure. Simulation studies demonstrate the satisfactory finite sample performance of the proposed method. Finally, our model is applied to the studies of lung cancer and type 2 diabetic complications.

Session 23INT8: High-Dimensional Data Analysis: Classification and Testing

A Powerful Methodology for Analyzing Correlated High Dimensional Data with Factor Models

Peng Wang[1], *Pengfei Lyu*[2], *Shyamal Peddada*[3] *and* ♦*Hongyuan Cao*[2]

[1]Jilin University
[2]Florida State University
[3]NIEHS

Multiple testing under dependence is a fundamental problem in high-dimensional statistical inference. We use a factor model to capture the dependence. Existing literature with factor models impose joint normality on the data or require tuning parameters to obtain robust inference. In this paper, we look at the problem from a different perspective by transposing approximate factor models. This allows heteroscedasticity and a more accurate estimation of the covariance matrix of idiosyncratic errors by projections. We construct factor-adjusted one-sample and two-sample test statistics of high-dimensional data. Extensive simulation studies demonstrate favorable performance of the proposed method over state-of-the-art methods while controlling the false discovery rate, even for heavy-tailed data. The robustness and tuning parameter-free features make the proposed method attractive to practitioners.

Multi-Threshold Structural Equation Model

Jingli Wang
Nankai University

In this paper, we consider the instrumental variable estimation for causal regression parameters with multiple unknown structural changes across subpopulations. We propose a multiple change point detection method to determine the number of thresholds and estimate the threshold locations in the two-stage least squares procedure. After identifying the estimated threshold locations, we use the Wald method to estimate the parameters of interest, i.e., the regression coefficients of the endogenous variable. Based on some technical assumptions, we carefully establish the consistency of estimated parameters and the asymptotic normality of causal coefficients. Simulation studies are included to examine the performance of the proposed method. Finally, our method is illustrated via an application of the Philippine farm households data for which some new findings are discovered.

Empirical Likelihood Ratio Tests for Nonnested Model Selection Based on Predictive Losses

Jiancheng Jiang[1], *Xuejun Jiang*[2] *and* ♦*Haofeng Wang*[2]

[1]University of North Carolina at Charlotte
[2]Southern University of Science and Technology

We propose an empirical likelihood ratio (ELR) test for comparing any two supervised learning models, where the competing models may be nested, nonnested, overlapping, misspecified, or correctly specified. It compares the prediction losses of models based on the cross-validation. We develop its asymptotic distributions for comparing two nonparametric learning models under a general framework with convex loss functions. However, the prediction losses from the cross-validation involve repeatedly fitting the models with one observation left out, which is a heavy computational burden. We introduce an easy-to-implement ELR test which requires fitting the models only once

and shares the same asymptotics as the original one. The proposed tests are applied to compare additive models with varying-coefficient models. Furthermore, a scalable distributed ELR test is proposed for testing the importance of a group of variables in possibly misspecified additive models with massive data. Simulations show that the proposed tests work well and have favorable finite sample performance over some existing approaches. The methodology is validated on an empirical application.

Asymptotic Normality for Eigenvalue Statistics of a General Sample Covariance Matrix when p/n →∞ and Applications

Jiaxin Qiu[1], ♦Zeng Li[2] and Jianfeng Yao[3]
[1]The University of Hong Kong
[2]Southern University of Science and Technology
[3]Chinese University of Hong Kong, Shenzhen

The asymptotic normality for a large family of eigenvalue statistics of a general sample covariance matrix is derived under the ultra-high dimensional setting, that is, when the dimension to sample size ratio $p/n \to \infty$. Based on this CLT result, we extend the covariance matrix test problem to the new ultrahigh dimensional context, and apply it to test a matrix-valued white noise. Simulation experiments are conducted for the investigation of finite-sample properties of the general asymptotic normality of eigenvalue statistics, as well as the two developed tests.

Session 23INT71: Recent Advances in Functional Data Analysis

Testing Linearity in Semi-Functional Partially Linear Regression Models

♦Yongzhen Feng[1], Jie Li[2] and Xiaojun Song[3]
[1]Tsinghua University
[2]Renmin University of China
[3]Peking University

This paper proposes a Kolmogorov–Smirnov type statistic and a Cramer–von Mises type statistic to test linearity in semi-functional partially linear regression models. Our test statistics are based on a residual marked empirical process indexed by a randomly projected functional covariate, which is able to circumvent the "curse of dimensionality" brought by the functional covariate. The asymptotic properties of the proposed test statistics under the null, the fixed alternative, and a sequence of local alternatives converging to the null at the n1/2 rate are established. A straightforward wild bootstrap procedure is suggested to estimate the critical values that are required to carry out the tests in practical applications. Results from an extensive simulation study show that our tests perform reasonably well in finite samples. Finally, we apply our tests to the Tecator and AEMET datasets to check whether the assumption of linearity is supported by these datasets.

Time-Varying Treatment Effects of Functional Data with Latent Confounders: Application to Sleep Heart Health Studies

♦Jie Li[1], Shujie Ma[2] and Yehua Li[2]
[1]Renmin University of China
[2]University of California, Riverside

Exploring the causal effect between variables is an important issue in lots of scientific research. Existing literature on causal inference mainly studies one-dimensional or multi-dimensional data, but functional data with repeated observations per individual frequently appears in a wide variety of applications. In functional data, treatment design may change across time, and its treatment effect is a time-varying function. Besides, most methods for treatment effect estimation based on observational data rely on the ignorability assumption that treatment assignment is independent of the potential outcomes given the observable covariates. This assumption can be violated when unobserved latent covariates are involved. We propose a novel method for unbiased treatment effect estimation with unobserved latent covariates for functional data. We propose to solve this challenging problem using a joint likelihood method with a Monte Carlo EM algorithm. Moreover, our proposed method is flexible to estimate both heterogeneous treatment effect of individuals and average treatment effect, providing a reliable inferential tool in making treatment decisions. It can also be applied to the irregular and sparse data. The method leads to meaningful discoveries when applied to investigate the dynamic effect of sleep quality on heart rate variability.

Ftir Feature Extraction with Forensic Microtraces Combining Screening and Functional Discriminant Analysis

Xing Wang
School of Statistics, Renmin University of China

Microtraces have the ability to move between the objects used in criminal activity. By establishing the chain of evidence and the microchemical structure, FTIR is a widely used technology that aids forensic scientists in the identification of evidence. Most current approaches are designed for marginal dimention reduction but ignore the effective range estimation due to the high dimention and complex chemical materials between microtraces. The chosen features are therefore not accurate enough. In this talk, we present a novel method for analyzing various types of microtraces that combines screening techniques with functional discriminant analysis. In order to make a discrimination, we collected 41 extremely similar samples from electric bike handlebar grip materials. We then used Fourier transform infrared spectroscopy to create the database (FTIR). The present work demonstrates the efffectives of the method we proposed for FTIR feature extraction as well as a scientific basis for forensic examination of microtraces. The current work provides a scientific foundation for forensic analysis of microtraces as well as proof of the effectiveness of the technique we suggested for FTIR feature extraction.

Network Vector Autoregression with Time-Varying Nodal Influence

♦Yi Ding[1], Rui Pan[2] and Bo Zhang
[1]Renmin University of China
[2]Central University of Finance and Economics

Vector autoregressive (VAR) models are a widely used class of time series models that have received considerable attention in the literature. However, in high- dimensional settings where the number of variables (nodes) in the time series is relatively large, transition matrix in VAR models is difficult to estimate. By incorporating network structure into VAR models, the number of parameters can be significantly reduced. In this paper we propose a time-varying network vector autoregressive (tvNAR) model. In the tvNAR model, the response of each node at a given time point is assumed to be the linear combination of previous values of its own and its connected neighbors in the network. The coefficients are assumed to be node-specific and

time-varying. The tvNAR model can reflect the unique effect of each node and is capable of describing the behavior of non-stationary time series. We propose a locally linear regression estimator of the time-varying nodal coefficients and establish its asymptotic properties. In order to examine the temporal constancy of the coefficients, we propose a Wald- type test. The performance of the estimator along with the test procedure is demonstrated through simulation studies and an empirical example of Nasdaq daily stock prices data.

Session 23INT13: Recent Developments on Causal Inference and Genetics

Staarpipeline: An all-in-One Rare-Variant Tool for Biobank-Scale Whole-Genome Sequencing Data

♦Zilin Li[1], Xihao Li[2] and Xihong Lin[2]
[1]Indiana University School of Medicine
[2]Harvard T.H. Chan School of Public Health

Large-scale whole-genome sequencing (WGS) studies have enabled the analysis of rare variant associations with complex human diseases and traits. Variant set analysis is a powerful approach to studying rare variant associations. However, existing methods have limited ability to define the variant set in the genome, especially for the noncoding genome. We propose a computationally efficient and robust rare variant association-detection framework, STAARpipeline, to automatically annotate a WGS study and perform flexible rare variant association analysis, including gene-centric analysis and fixed-window and dynamic-window-based non-gene-centric analysis by incorporating variant functional annotations. In gene-centric analysis, STAARpipeline groups coding and noncoding variants based on functional categories of genes and incorporate multiple functional annotations. In non-gene-centric analysis, in addition to fixed-size sliding window analysis, STAARpipeline provides a data-adaptive-size dynamic window analysis. All these variant sets could be automatically defined and selected in STAARpipeline. STAARpipeline also provides analytical follow-up of dissecting association signals independent of known variants via conditional analysis. We applied the STAARpipeline to analyze the total cholesterol in 30,138 samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. All analyses scale well in computation time and memory. We discover several potentially new significant associations with lipids, including a finding of rare variants in an intergenic region near JKAMPP1 associated with total cholesterol. In summary, the STAARpipeline is a powerful and resource-efficient tool for association analysis of biobank-scale WGS studies.

Causal Inference with Invalid Instruments: Post-Selection Problems and a Solution using Searching and Sampling

Zijian Guo
Rutgers

Instrumental variable methods are among the most commonly used causal inference approaches to deal with unmeasured confounders in observational studies. The presence of invalid instruments is the primary concern for practical applications, and a fast-growing area of research is inference for the causal effect with possibly invalid instruments. This paper illustrates that the existing confidence intervals may undercover when the valid and invalid instruments are hard to separate in a data-dependent way. To address this, we construct uniformly valid confidence intervals that are robust to the mistakes in separating valid and invalid instruments. We propose to search for a range of treatment effect values that lead to sufficiently many valid instruments. We further devise a novel sampling method, which, together with searching, leads to a more precise confidence interval. Our proposed searching and sampling confidence intervals are uniformly valid and achieve the parametric length under the finite-sample majority and plurality rules. We apply our proposal to examine the effect of education on earnings. The proposed method is implemented in the R package $\mathrm{RobustIV}$ available from CRAN.

Optimal Individualized Decision-Making with Proxies

Tao Shen[1] and ♦Yifan Cui[2]
[1]NUS
[2]ZJU

A common concern when a policymaker draws causal inferences from and makes decisions based on observational data is that the measured covariates are insufficiently rich to account for all sources of confounding, i.e., the standard no confoundedness assumption fails to hold. The recently proposed proximal causal inference framework shows that proxy variables can be leveraged to identify causal effects and therefore facilitate decision-making. Building upon this line of work, we propose a novel optimal individualized treatment regime based on so-called outcome-inducing and treatment-inducing confounding bridges. We then show that the value function of this new optimal treatment regime is superior to that of existing ones in the literature. Theoretical guarantees, including identification, superiority, and excess value bound of the estimated regime, are established. Furthermore, we demonstrate the proposed optimal regime via numerical experiments and a real data application.

Recent Progress in Machine Learning-Based Causal Inference

Lin Liu
Shanghai Jiao Tong University

In this talk, we will discuss some of our recent works in causal effect estimation, with a particular emphasis on doubly robust functionals. An important research problem is to characterize the sufficient and necessary conditions under which causal effects can be estimated at root-n-rate. We consider this problem in three different scenarios – (1) the nonparametric scenario, (2) the high-dimensional sparsity scenario, and (3) what happens if we use deep neural networks. For (1) and (2), we present the minimal conditions for root-n-rate and estimators with good practical performance. For (3), we present the theoretical result we can achieve based on the existing tools, show its limitation, and discuss some future directions.

Session 23INT31: New Development on Precision Medicine

Dynamic Logistic State Space Prediction Model for Clinical Decision Making

♦Jiakun Jiang[1], Wei Yang[2], Erin M. Schnellinger[2], Stephen E. Kimmel[2] and Wensheng Guo[2]
[1]Beijing Normal University
[2]University of Pennsylvania

Prediction modeling for clinical decision making is of great importance and needed to be updated frequently with the changes

of patient population and clinical practice. Existing methods are either done in an ad hoc fashion, such as model recalibration or focus on studying the relationship between predictors and outcome and less so for the purpose of prediction. In this article, we pro- pose a dynamic logistic state space model to continuously update the parameters whenever new information becomes available. The proposed model allows for both time-varying and time-invariant coefficients. The varying coefficients are modeled using smoothing splines to account for their smooth trends over time. The smoothing parameters are objectively chosen by maximum likelihood. The model is updated using batch data accumulated at prespecified time intervals, which allows for better approximation of the underlying binomial density func- tion. In the simulation, we show that the new model has significantly higher pre- diction accuracy compared to existing methods. We apply the method to predict 1 year survival after lung transplantation using the United Network for Organ Sharing data.

Center-Augmented L_2-Type Regularization for Subgroup Learning

♦Ye He[1], Ling Zhou[2], Yingcun Xia[3] and Huazhen Lin[2]
[1]Sichuan Normal University,China
[2]Southwestern University of Finance and Economics,China
[3]National University of Singapore, Singapore

The existing methods for subgroup analysis can be roughly divided into two categories: finite mixture models (FMM) and regularization methods with an l_1-type penalty. In this paper, by introducing the group centers and l_2-type penalty in the loss function, we propose a novel center-augmented regularization(CAR) method; this method can be regarded as a unification of the regularization method and FMM and hence exhibits higher efficiency and robustness and simpler computations than the existing methods. In particular, its computational complexity is reduced from the $O(n2)$ of the conventional pairwise-penalty method to only $O(nK)$, where n is the sample size and K is the number of subgroups. The asymptotic normality of CAR is established, and the convergence of the algorithm is proven. CAR is applied to a dataset from a multicenter clinical trial, Buprenorphine in the Treatment of Opiate Dependence; a larger R2 is produced and three additional significant variables are identified compared to those of the existing methods.

Generalized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types

♦Wei Liu[1], Huazhen Lin[2], Shurong Zheng[3] and Jin Liu[4]
[1]Centre for Quantitative Medicine, Program in Health Services & Systems Research, Duke-NUS Medical School
[2]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics
[3]School of Mathematics and Statistics, Northeast Normal University
[4]School of Data Science, The Chinese University of Hong Kong-Shenzhen

As high-dimensional data measured with mixed-type variables gradually become prevalent, it is particularly appealing to represent those mixed-type high-dimensional data using a much smaller set of so-called factors. Due to the limitation of the existing methods for factor analysis that deal with only continuous variables, in this article, we develop a generalized factor model, a corresponding algorithm and theory for ultra-high dimensional mixed types of variables where both the sample size n and vari-

able dimension p could diverge to infnity. Specifcally, to solve the computational problem arising from the non-linearity and mixed types, we develop a two-step algorithm so that each update can be carried out in parallel across variables and samples by using an existing package. Theoretically, we establish the rate of convergence for the estimators of factors and loadings in the presence of nonlinear structure accompanied with mixed-type variables when both n and p diverge to infnity. Moreover, since the correct specifcation of the number of factors is crucial to both the theoretical and the empirical validity of factor models, we also develop a criterion based on a penalized loss to consistently estimate the number of factors under the framework of a generalized factor model. To demonstrate the advantages of the proposed method over the existing ones, we conducted extensive simulation studies and also applied it to the analysis of the NFBC1966 dataset and a cardiac arrhythmia dataset, resulting in more predictive and interpretable estimators for loadings and factors than the existing factor model.

Subgroup-Effects Models for the Analysis of Personal Treatment Effects

♦Ling Zhou[1], Shiquan Sun[2], Haoda Fu[3] and Peter Song[4]
[1]Southwestern University of Finance and Economics
[2]Xi'an Jiaotong University
[3]Eli Lilly and Company
[4]University of Michigan

The emerging field of precision medicine is transforming statistical anal- ysis from the classical paradigm of population-average treatment effects into that of personal treatment effects. This new scientific mission has called for adequate statistical methods to assess heterogeneous covariate effects in re- gression analysis. This paper focuses on a subgroup analysis that consists of two primary analytic tasks: identification of treatment effect subgroups and individual group memberships, and statistical inference on treatment effects by subgroup. We propose an approach to synergizing supervised clustering analysis via Alternating Direction Method of Multipliers (ADMM) algorithm and statistical inference on subgroup effects via Expectation-Maximization (EM) algorithm. Our proposed procedure, termed as Hybrid Operation for Subgroup Analysis (HOSA), enjoys computational speed and numerical sta- bility with interpretability and reproducibility. We establish key theoretical properties for both proposed clustering and inference procedures. Numerical illustration includes extensive simulation studies and analyses of motivating data from two randomized clinical trials to learn subgroup treatment effects.

Session 23INT18: High-Dimensional Regression, State Space Models, and COVID-19 Prediction

Sparse Convoluted Rank Regression in High Dimensions

♦Le Zhou[1], Boxiang Wang[2] and Hui Zou[3]
[1]Hong Kong Baptist University
[2]University of Iowa
[3]University of Minnesota

Wang et al. (2020, JASA) studied the high-dimensional sparse penalized rank regression and established its nice theoretical properties. Compared with the least squares, rank regression can have a substantial gain in estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the

computation of penalized rank regression can be very challenging for high-dimensional data, due to the highly nonsmooth rank regression loss. In this work we view the rank regression loss as a non-smooth empirical counterpart of a population level quantity, and a smooth empirical counterpart is derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-dimensional data. We prove some interesting asymptotic properties of CRR. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the ℓ_1-penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. We derive an efficient algorithm for solving sparse convoluted rank regression that scales well with high dimensions. Numerical examples demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression. Our theoretical and numerical results suggest that sparse convoluted rank regression enjoys the best of both sparse least squares regression and sparse rank regression.

Dynamic Hierarchical State Space Forecasting

♦*Ziyue Liu*[1] and *Wensheng Guo*[2]
[1]Indiana University School of Medicine
[2]University of Pennsylvania

We propose a new prediction and forecast framework using hierarchical state space models. This approach borrows information on estimating both the mean processes shared by all the subjects and the parameters governing subject-specific dynamics. Application to the state level COVID-19 case numbers in USA demonstrates that the proposed approach outperforms classical prediction models.

Predicting Sars-Cov-2 Infection among Hemodialysis Patients using Multimodal Data

♦*Juntao Duan*[1], *Hanmo Li*[1], *Xiaoran Ma*[1], *Yuedong Wang*[1], *Peter Kotanko*[2], *Hanjie Zhang*[3], *Rachel Lasky*[4], *Caitlin Monaghan*[4], *Mengyang Gu*[1] and *Wensheng Guo*[5]
[1]Department of Statistics and Applied Probability, University of California, Santa Barbara, California, United States
[2]Icahn School of Medicine at Mount Sinai, New York, United States
[3]Renal Research Institute, New York, United States
[4]Fresenius Medical Care, Global Medical Office, Waltham, Massachusetts
[5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States

The COVID-19 pandemic has created more devastation to dialysis patients than to the general population. Patient-level prediction models for SARS-CoV-2 infection are crucial for the early identification of patients to prevent and mitigate outbreaks within dialysis clinics. As the COVID-19 pandemic evolves, it is unclear whether previously built prediction models are still sufficiently effective. We developed a machine learning (XGBoost) model to predict during the incubation period a SARS-CoV-2 infection that is subsequently diagnosed after three or more days. We used data from multiple sources, including demo-graphic, clinical, treatment, laboratory, and vaccination information from a national network of hemodialysis clinics, socioeconomic information from the Census Bureau, and county-level COVID-19 infection and mortality information from state and local health agencies. We created prediction models and evaluated their performances on a rolling basis to investigate the evolution of prediction power and risk factors.

Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with Covid-19

Xiaoran Ma[1], *Wensheng Guo*[2], *Mengyang Gu*[1], *Peter Kotanko*[3], *Len Usvyat*[4] and ♦*Yuedong Wang*[1]
[1]University of California - Santa Barbara
[2]University of Pennsylvania
[3]Renal Research Institute
[4]Fresenius Medical Care

Some patients with COVID-19 show changes in signs and symptoms, such as temperature and oxygen saturation, days before being positively tested for SARS-CoV-2, while others remain asymptomatic. It is important to identify these subgroups and to understand what biological and clinical predictors are related to these subgroups. This information will provide insights into how the immune system may respond differently to infection and can further be used to identify infected individuals. We propose a flexible nonparametric mixed-effects mixture model that identifies risk factors and classifies patients with biological changes. We model the latent probability of biological changes using a logistic regression model and trajectories in the latent groups using smoothing splines. We developed an EM algorithm to maximize the penalized likelihood for estimating all parameters and mean functions. We evaluate our methods by simulations and apply the proposed model to investigate changes in temperature in a cohort of COVID-19-infected hemodialysis patients.

Session 23INT12: Recent Developments in Analysis of Functional, Longitudinal, and Time-to-Event Data

Functional Data Analysis with Covariate-Dependent Mean and Covariance Structures*

Huazhen Lin
Southwestern University of Finance and Economics

Functional data analysis has emerged as a powerful tool in response to the ever increasing resources and efforts devoted to collecting information about response curves or anything varying over a continuum. However, limited progress has been made to link the covariance structure of response curves to external covariates, as most functional models assume a common covariance structure. We propose a new functional regression model with covariate-dependent mean and covariance structures. Particularly, by allowing the variances of the random scores to be covariate-dependent, we identify eigenfunctions for each individual from the set of eigenfunctions which govern the patterns of variation across all individuals, resulting in high interpretability and prediction power. We further propose a new penalized quasi-likelihood procedure, which combines regularization and B-spline smoothing, for model selection and estimation, and establish the convergence rate and asymptotic normality for the proposed estimators. The utility of the method is demonstrated via simulations as well as an analysis of the Avon Longitudinal Study of Parents and Children on parental effects on the growth

curves of their offspring, which yields biologically interesting results.

Robust Inference for Joint Models of Longitudinal and Survival Data

Lang Wu

University of British Columbia, Vancouver

In a survival model with a time-dependent covariate, the covariate may be left censored due to a lower detection limit and its observed values may contain outliers. Motivated from an HIV vaccine study, we propose a robust method for joint models of longitudinal and survival data, where the outliers in longitudinal data are addressed using a multivariate t-distribution for b-outliers and using an M-estimator for e-outliers. We also propose a computationally efficient method for approximate likelihood inference. The proposed method is evaluated by simulation studies. Based on the proposed models and method, we analyze the HIV vaccine data and find a strong association between longitudinal biomarkers and the risk of HIV infection.

Functional Data Modeling in High Dimensions: Fundamentals, Sparsity and Fast Computation.

Shaojun Guo

Covariance function estimation is a fundamental task in multivariate functional data analysis and arises in many applications. In this paper, we consider estimating sparse covariance functions for high-dimensional functional data, where the number of random functions p is comparable to, or even larger than the sample size n. Aided by the Hilbert-Schmidt norm of functions, we introduce a new class of functional thresholding operators that combine functional versions of thresholding and shrinkage, and propose the adaptive functional thresholding estimator by incorporating the variance effects of individual entries of the sample covariance function into functional thresholding. To handle the practical scenario where curves are partially observed with errors, we also develop a nonparametric smoothing approach to obtain smoothed adaptive functional thresholding estimator and its binned implementation to accelerate the computation. We investigate the theoretical properties of our estimators when p grows exponentially with n under both fully and partially observed functional scenarios. Finally, we demonstrate that the proposed adaptive functional thresholding estimators significantly outperform the competing estimators through extensive simulations and the functional connectivity analysis of two neuroimaging datasets.

Session 23INT94: Advanced Statistical Methods for Complex Observational Studies and Clinical Trials

Survival Analysis of Randomized Controlled Trials with Auxiliary Patient Population Information from Observational Studies

Xiaofei Wang

Duke University

In the presence of heterogeneity between the randomized controlled trial (RCT) participants and the target population, evaluating the treatment effect solely based on the RCT often leads to biased quantification of the real-world treatment effect. To address the problem of lack of generalizability for the treatment effect estimated by the RCT sample, we leverage observational studies with large samples that are representative of the target population. The focus is on evaluating treatment effects on survival outcomes for a target population. A broad class of estimands is considered that are functionals of treatment-specific survival functions, including differences in survival probability and restricted mean survival times. We propose a semiparametric estimator through the guidance of the efficient influence function. The proposed estimator is doubly robust in the sense that it is consistent for the target population estimands if either the survival model or the weighting model is correctly specified, and is locally efficient when both are correct. Simulation studies confirm the theoretical properties of the proposed estimator and show it outperforms competitors. We apply the proposed method to estimate the effect of adjuvant chemotherapy on survival in patients with early-stage resected non-small lung cancer.

Subgroup Analysis for Longitudinal Data Based on a Partial Linear Varying Coefficient Model with a Change Plane

Guoyou Qin

Fudan University

Subgroup analysis has become an important tool to characterize the treatment effect heterogeneity, and finally towards precision medicine. On the other hand, longitudinal study is widespread in many fields, but subgroup analysis for this data type is still limited. In this paper, we study a partial linear varying coefficient model with a change plane, in which the subgroups are defined based on linear combination of grouping variables, and the time-varying effects in different subgroups are estimated to capture the dynamic association between predictors and response. The varying coefficients are approximated by basis functions and the group indicator function is smoothed by kernel function, which are included in the generalized estimating equation for estimation. Asymptotic properties of the estimators for the varying coefficients, the constant coefficients and the change plane coefficients are established. Simulations are conducted to demonstrate the flexibility, efficiency and robustness of the proposed method. Based on the Standard and New Antiepileptic Drugs study, we successfully identify a subgroup in which patients are sensitive to the newer drug in a specific period of time.

Practical Considerations in Trial Design with Win Ratio Method for Multiple Time-to-Event Endpoints with Hierarchy

♦*Huiman Barnhart, Yuliya Lokhnygina, Roland Matsouaka and Frank Rockhold*

Duke University

Standard approach to the analysis of composite endpoints is time-to-first event analysis. However, this approach has been criticized because it ignores the differences in clinical severity or importance and may end up emphasizing the less severe or non-terminal events while ignoring the severe events occurring later. Finkelstein and Schoenfeld (1999) considered the use of hierarchical endpoints to take advantage of the ranking that is based on the severity of the component endpoints. Pocock et al. (2012) popularized the win ratio method that takes advantage of clinical priority in multiple endpoints while providing an estimate of treatment effect that is interpretable clinically. Redfors et al. (2020) and Ferreira et al. (2020) re-analyzed several trials using the win ratio method and found that the results with win ratio yielded similar conclusions as those from the original standard time-to-first event analysis. As the win ratio is becoming more popular due to its intuitive appeal, design of

randomized clinical trials using the win ratio is lagging behind. We address two specific practical issues. We first examine the expected magnitude of win ratio for time to event endpoints with hierarchy. Second, we explore scenarios where the win ratio method has greater statistical power than time-to-first event analysis via in-depth simulations. Several examples are used to illustrate these two issues. We provide guidance on when to use win ratio in designing a study with hierarchical endpoints and on how to choose the magnitude of win ratio parameter for power calculations.

Accelerated Failure Time Modeling via Nonparametric Mixtures

Byungtae Seo[1] and ♦*Sangwook Kang*[2]
[1]Sungkyunkwan University
[2]Yonsei University

An accelerated failure time (AFT) model assuming a log-linear relationship between failure time and a set of covariates can be either parametric or semi-parametric, depending on the distributional assumption for the error term. Both classes of AFT models have been popular in the analysis of censored failure time data. The semiparametric AFT model is more flexible and robust to departures from the distributional assumption than its parametric counterpart. However, the semiparametric AFT model is subject to producing biased results for estimating any quantities involving an intercept. Estimating an intercept requires a separate procedure. Moreover, a consistent estimation of the intercept requires stringent conditions. Thus, essential quantities such as mean failure times might not be reliably estimated using semiparametric AFT models, which can be naturally done in the framework of parametric AFT models. Meanwhile, parametric AFT models can be severely impaired by misspecifications. To overcome this, we propose a new type of the AFT model using a nonparametric Gaussian-scale mixture distribution. We also provide feasible algorithms to estimate the parameters and mixing distribution. The finite sample properties of the proposed estimators are investigated via an extensive stimulation study. The proposed estimators are illustrated using the well-known PBC dataset.

Session 23INT19: Advanced Statistical Learning Methods for Heterogeneous Data and Model Integration

Evaluation of Combined Data from Subgroup Selection and Validation Phases in Clinical Trials

♦*Xinzhou Guo*[1], *Jianjun Zhou*[2] and *Xuming He*[3]
[1]Hong Kong University of Science and Technology
[2]Yunnan University
[3]University of Michigan

When a promising subgroup is identified from an unsuccessful trial with a broad target population, we often need to evaluate and possibly confirm the selected subgroup with a follow-up validation trial. A direct evaluation of the subgroup from the subjects in both trials is not recommended because of the risk of data snooping. An evaluation based solely on the validation trial is free of bias, but does not make full use of the data in the earlier trial. We show that it is possible to utilize data from both trials to improve the efficiency of post-selection subgroup evaluation. In particular, we propose a new resampling-based

approach to quantify and remove selection bias and then to perform data combination from both trials for valid and efficient inference on selected subgroup. The proposed method is model-free and asymptotically sharp. We demonstrate the merit of the proposed method by revisiting the panitumumab trial and show how much data combination could help improve efficiency of clinical trials when a promising subgroup is identified post hoc from part of the data.

d-Gcca: Decomposition-Based Generalized Canonical Correlation Analysis for Multi-View High-Dimensional Data

♦*Hai Shu*[1], *Zhe Qu*[2] and *Hongtu Zhu*[3]
[1]New York University
[2]Tulane University
[3]The University of North Carolina at Chapel Hill

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A popular model in high-dimensional multi-view data analysis is to decompose each view's data matrix into a low-rank common-source matrix generated by latent factors common across all data views, a low- rank distinctive-source matrix corresponding to each view, and an additive noise matrix. We propose a novel decomposition method for this model, called decomposition-based generalized canonical correlation analysis (D-GCCA). The D-GCCA rigorously defines the decomposition on the L2 space of random variables in contrast to the Euclidean dot product space used by most existing methods, thereby being able to provide the estimation consistency for the low-rank matrix recovery. Moreover, to well calibrate common latent factors, we impose a desirable orthogonality constraint on distinctive latent factors. Existing methods, however, inadequately consider such orthogonality and may thus suffer from substantial loss of undetected common-source variation. Our D-GCCA takes one step further than generalized canonical correlation analysis by separating common and distinctive components among canonical variables, while enjoying an appealing interpretation from the perspective of principal component analysis. Furthermore, we propose to use the variable-level proportion of signal variance explained by common or distinctive latent factors for selecting the variables most influenced. Consistent estimators of our D-GCCA method are established with good finite-sample numerical performance, and have closed-form expressions leading to efficient computation especially for large-scale data. The superiority of D-GCCA over state-of-the-art methods is also corroborated in simulations and real-world data examples.

Robust Transfer Learning of Individualized Treatment Rules

Lu Tang
University of Pittsburgh

Causality-based individualized treatment rules (ITRs) are a stepping-stone to precision medicine. To ensure unconfoundedness, ITRs are ideally derived from randomized experimental data, but the use cases of ITRs in the real-world data extend far beyond these controlled settings. It is of great interest to transfer knowledge learned from experimental data to real-world data, but hurdles remain. In this paper, we address two challenges in the transfer learning of ITRs. 1) In well-designed experiments, granular information crucial to decision making can be thoroughly collected. However, part of this may not be accessible in real-world decision-making. 2) Experimental data with strict inclusion criteria reflect a population distribution that may be very different from the real-world population data, leading

to suboptimal ITRs. We propose a unified weighting scheme to learn a calibrated and robust ITR that simultaneously addresses the issues of covariate shift and missing covariates during prospective deployment, with a quantile-based approach to ensure worst-case safety under the uncertainty due to unavailable covariates. The performance of this method is evaluated in simulations and real-data applications.

Session 23INT67: Novel Statistical Models and Methods with Applications

Bayesian Inference and Applications for Zero-Inflated Models

Seong Kim
Hanyang University

Analysis of discrete data is frequently conducted in diverse fields, including natural sciences, social sciences, public health , and other disciplines. The binomial and Poisson distributions are perhaps commonly used in utilizing discrete data. For instance, it would be interesting to check the number of defective products in total production; to observe how many earthquakes will occur in one year; or to see how many home runs can be produced by a baseball batter in each game. More often than not, these count data possess a considerable number of excessive zeros, hindering analysis with the regular binomial and Poisson distributions. Under these circumstances with excessive zero patterns, zero-inflated models would be a remedy to circumvent loss of information or tendencies of biased estimators. In this talk, several zero-inflated models associated with binomial, bivariate binomial, and Poisson distributions are analyzed. Prior elicitation issues in conjunction with each model are presented. Several real datasets are analyzed to support theoretical results.

The Gender Gap in Venture Capital Market: a Statistical Approach using Structural Matching Models and Accelerator Data

Chuan Chen[1] *and* ♦*Junnan He*[2]
[1]Wisconsin School of Business
[2]Sciences Po

In this study, we utilize a two-sided structural matching model to examine the gender gap in the venture capital (VC) market. To account for unobservable startup qualities at the time of entry into the VC market, we construct a unique dataset comprising accelerator participants. Subsequently, we impute the unobserved quality by modeling the accelerator admission as a two-sided matching process. After adjusting for this unobserved quality, our analysis shows that female entrepreneurs have a lower probability of securing large venture capital investments. Our further analysis does not support potential explanations such as gender differences in risk aversion or relocation preferences. From a methodological perspective, we introduce an efficient and accessible estimation algorithm for the Sorensen (2007) model, applicable to various research contexts.

Semiparametric Evaluation of First-Passage Distribution for Step-Stress Accelerated Degradation Tests

Lochana Palayangoda[1]*,* ♦*Hon Keung Tony Ng*[2] *and Ling Li*[3]
[1]Department of Mathematical and Statistical Sciences, University of Nebraska Omaha
[2]Department of Mathematical Sciences, Bentley University
[3]Xi'an Microelectronic Technology Institute

In reliability engineering, different types of accelerated degradation tests have been used to obtain information for evaluating highly reliable or expensive products. Step-stress accelerated degradation test (SSADT) is one of the useful experimental schemes that can be used to save the resources of an experiment. Motivated by the SSADT data for operational amplifiers collected in Xi'an Microelectronic Technology Institute, in which the underlying degradation mechanism of the operational amplifiers is unknown, we propose a semiparametric approach for SSADT data analysis that does not require strict distributional assumptions. Specifically, the empirical saddlepoint approximation method is proposed to estimate the items' lifetime (first-passage time) distribution at both stress levels included and not included in the SSADT experiment. Monte Carlo simulation studies are used to evaluate the performance and illustrate the advantages of the proposed approach. Finally, the proposed semiparametric approach is applied to analyze the motivating data set.

Session 23INT58: Statistical Genetics and Genomics

Transfer Learning for High-Dimensional Multiple Response Regression

Seyoung Park
Sungkyunkwan University

In this talk, I will discuss the transfer learning problem under high-dimensional multiple response linear models, in which the underlying regression coefficient matrix has low-rank. Transfer learning aims to improve the fit on the target data by borrowing auxiliary samples from different but possibly related source data. When the set of informative sources is known, we propose a transfer learning algorithm based on nuclear norm penalization, and derive its estimation and prediction error bounds, which is faster than the rates without using the informative source data. When the set of informative sources is unknown, a source detection approach is proposed to detect informative sources and the detection consistency is proved. In the analysis of cancer patient data consisting of multiple cancer subtypes, the proposed method leads to improved performance in gene expression prediction in a target cancer subtype by incorporating the data from different cancer subtypes.

Decoding Gene Functions: Exploring their Significance in Biological Context

Ying Zhu
Fudan University

The prevalence of high-throughput technologies has greatly accelerated the discovery of genes with specific functions and their associations with diseases. However, the challenge remains in mechanistic interpretation of these discoveries. Here, we introduce the bioinformatics tools we developed that infer tissue and cell-specific gene functions.

Multi-Tissue Transcriptome-Wide Association Studies with High Dimensional Transfer Learning

♦*Daoyuan Lai, Han Wang and Yan Dora Zhang*
Department of Statistics and Actuarial Science, The University of Hong Kong

Transcriptome-wise association study (TWAS) is a novel approach to studying the association between a patient's genotype and corresponding complex trait. TWAS imputes gene expression levels from genotypes through samples with matched

genotypes and gene expression levels in a fixed human tissue; thus, the genetic architecture of complex traits is estimated by the imputed gene expression level. However, a major challenge is building a robust and accurate imputation model for tissues with a limited sample size. This paper introduced an efficient transfer learning algorithm to leverage information from external tissues to improve the prediction performance of a target model. The key advantage of the algorithm is we explicitly considered the concordance between external and target tissues; therefore, information from similar external tissues will contribute more to the information borrowing. A simulation and a real data analysis showed that our algorithm increases the imputation accuracy by 20% on average compared with existing methods. We also applied the algorithm to multiple genome-wide association results and showed improvements over existing single- or multiple-tissue methods.

A Statistical Framework for Cross-Population Fine-Mapping by Leveraging Genetic Diversity and Accounting for Confounding Bias

♦*Mingxuan Cai*[1], *Zhiwei Wang*[2], *Jiashun Xiao*[3], *Xianghong Hu*[2], *Gang Chen*[4] *and Can Yang*[2]
[1]City University of Hong Kong
[2]The Hong Kong University of Science and Technology
[3]Shenzhen Research Institute of Big Data
[4]The WeGene Company

Fine-mapping prioritizes risk variants identified by genome-wide association studies (GWASs), serving as a critical step to uncover biological mechanisms underlying complex traits. However, several major challenges still remain for existing fine-mapping methods. First, the strong linkage disequilibrium among variants can limit the statistical power and resolution of fine-mapping. Second, it is computationally expensive to simultaneously search for multiple causal variants. Third, the confounding bias hidden in GWAS summary statistics can produce spurious signals. To address these challenges, we develop a statistical method for cross-population fine-mapping (XMAP) by leveraging genetic diversity and accounting for confounding bias. By using cross-population GWAS summary statistics from global biobanks and genomic consortia, we show that XMAP can achieve greater statistical power, better control of false positive rate, and substantially higher computational efficiency for identifying multiple causal signals, compared to existing methods. Importantly, we show that the output of XMAP can be integrated with single-cell datasets, which greatly improves the interpretation of putative causal variants in their cellular context at single-cell resolution.

Session 23INT88: Recent Advances in Statistical Methods and Machine Learning.

Super-Quantile Regression

Xuming He[1], *Kean Ming Tan*[2] *and* ♦*Wenxin Zhou*[3]
[1]Washington University in St. Louis
[2]University of Michigan
[3]University of Illinois at Chicago

Expected Shortfall (ES), also known as super-quantile or Conditional Value-at-Risk, is an important measure in risk analysis and stochastic optimization and has applications beyond these fields. In finance, it refers to the conditional expected return of an asset given that the return is below some quantile of its distribution. In this talk, we consider a joint regression framework recently proposed to model the quantile and ES of a response variable simultaneously, given a set of covariates. The current state-of-the-art approach to this problem involves minimizing a non-differentiable and non-convex joint loss function, which poses numerical challenges and limits its applicability to large-scale data. Motivated by the idea of using Neyman-orthogonal scores to reduce sensitivity to nuisance parameters, we propose a statistically robust and computationally efficient two-step procedure for fitting joint quantile and ES regression models that can handle highly skewed and heavy-tailed data. We establish explicit non-asymptotic bounds on estimation and Gaussian approximation errors that lay the foundation for statistical inference, even with increasing covariate dimensions.

Complexity of Feed-Forward Neural Networks from the Perspective of Functional Equivalence

Guohao Shen
The Hong Kong Polytechnic University

In this paper, we investigate the complexity of feed-forward neural networks by examining the concept of functional equivalence, which suggests that different network parameterizations can lead to the same function. We utilize the permutation invariance property to derive a novel covering number bound for the class of feedforward neural networks, which reveals that the complexity of a neural network can be reduced by exploiting this property. Furthermore, based on the symmetric structure of parameter space, we demonstrate that an appropriate strategy of random parameter initialization can increase the probability of convergence for optimization. We found that overparameterized networks tend to be easier to train in the sense that increasing the width of neural networks leads to a vanishing volume of the effective parameter space. Our findings offer new insights into overparameterization and have significant implications for understanding generalization and optimization in deep learning.

Session 23INT20: Statistical Methods and Applications in Precision Medicine

Wemics: a Single-Base Resolution Methylation Quantification Method by Weighting Methylation of Consecutive Cpg Sites

Yi Liu
Zhejiang University

DNA methylation, an epigenetic mechanism that alters gene expression without changing DNA sequence, is essential for organism development and key biological processes like genomic imprinting and X-chromosome inactivation. Despite tremendous efforts in DNA methylation research, accurate quantification of cytosine methylation remains a challenge. Here, we introduced a single-base methylation quantification approach by weighting co-methylation of consecutive CpG sites (Wemics) in genomic regions. Wemics quantification of DNA methylation better predicts its regulatory impact on gene transcription and identifies differentially methylated regions (DMRs) with more biological relevance. Most Wemics-quantified DMRs in lung cancer were conserved and recurrently occurred in other primary cancers from The Cancer Genome Atlas (TCGA), and their aberrant alterations could serve as promising pan-cancer diagnostic markers. We further revealed that these detected

DMRs are enriched in transcription factor (TF) binding motifs, and methylation of these TF binding motifs and TF expression synergistically regulate target gene expression. Using Wemics on epigenomic-transcriptomic data from our large lung cancer cohort, we discovered a dozen novel genes with oncogenic potential that were upregulated by hypomethylation but overlooked by other quantification methods. These findings increase our understanding of the epigenetic mechanism by which DNA methylation regulates gene expression.

Robust Method for Optimal Treatment Decision Making Based on Survival Data

◆Yuexin Fang[1], Baqun Zhang[2] and Min Zhang[3]
[1]Shanghai Normal University
[2]Shanghai University of Finance and Economics
[3]University of Michigan

Identifying the optimal treatment decision rule, where the best treatment for an individual varies according to his/her characteristics, is of great importance when treatment effect heterogeneity exists. We develop methods for estimating the optimal treatment decision rule based on data with survival time as the primary endpoint. Our methods are based on a flexible semiparametric accelerated failure time model, where only the treatment contrast (i.e., the difference in means between treatments) is parameterized and all other aspects are unspecified. An individual's treatment contrast is firstly estimated robustly by an augmented inverse probability weighted estimator (AIPWE). Then the optimal decision rule is estimated by minimizing the loss between the treatment contrast and the AIPWE contrast. Two loss functions with different strategies to account for censoring are proposed. The proposed loss functions distinguish from existing ones in that they are based on treatment contrasts, which completely determine the optimal treatment rule. Our methods can further incorporate a penalty term to select variables that are only important for treatment decision making, while taking advantage of all covariates predictive of outcomes to improve performance. Comprehensive simulation studies have been conducted to evaluate performances of the proposed methods relative to existing methods. The proposed methods are illustrated with an application to the ACTG 175 clinical trial on HIV-infected patients.

Methods for Identifying Differentially Methylated Regions for Monozygotic Twins

◆Xiaoqing Pan[1], Pengyuan Liu[2], Srividya Kidambi[3] and Mingyu Liang[3]
[1]Shanghai Normal University
[2]Zhejiang University
[3]Medical College of Wisconsin

DNA methylation plays a vital role in gene transcriptional regulation. With the advent of next-generation sequencing technologies, reduced representation bisulfite sequencing (RRBS) is becoming increasingly common for analyzing genome-wide methylation profiles at the single nucleotide level. A major goal of RRBS studies is to detect differentially methylated regions (DMRs) between different biological conditions. Monozygotic twins, treated as unordered pairs, are classical epidemiological designs to examine the genetic and environmental influence in complex diseases. However, no DMR identification tool for paired samples is currently available. In our first study, we present an innovative computational tool, PMat, combing folded normal test with a binary segmentation algorithm, to identify DMRs in twin samples. In our second study, we provide a Gibbs sampler which has advantages in DNA methylation imputation for Monozygotic twins.

Driverwmds: a Powerful Network Control Method for Predicting Cancer Driver Genes

Xiang Cheng[1], Xiaoqing Pan[2], Pengyuan Liu[1] and ◆Yan Lu[1]
[1]Zhejiang University
[2]Shanghai Normal University

Mammalian cells can be transcriptionally reprogramed to other cellular phenotypes. Controllability of such complex transitions in transcriptional networks underlying cellular phenotypes is an inherent biological characteristic. This network controllability can be interpreted by operating a few key regulators to guide the transcriptional program from one state to another. Finding the key regulators in the transcriptional program can provide key insights into the network state transition underlying cellular phenotypes. To address this challenge, here, we proposed to identify the key regulators in the transcriptional co-expression network as a minimum dominating set (MDS) of driver nodes that can fully control the network state transition. Based on the theory of structural controllability, we developed a weighted MDS network model (WMDS.net) to find the driver nodes of differential gene co-expression networks. The weight of WMDS.net integrates the degree of nodes in the network and the significance of gene co-expression difference between two physiological states into the measurement of node controllability of the transcriptional network. To confirm its validity, we applied WMDS.net to the discovery of cancer driver genes in RNA-seq datasets from The Cancer Genome Atlas. WMDS.net is powerful among various cancer datasets and outperformed the other top-tier tools with a better balance between precision and recall. Availability and implementation: https://github.com/chaofen123/WMDS.net.

Session 23INT27: Inference for High Dimensional Data

Cp Factor Model for Dynamic Tensors

◆Yuefeng Han[1], Cun-Hui Zhang[2] and Rong Chen[2]
[1]University of Notre Dame
[2]Rutgers University

Observations in various applications are frequently represented as a time series of multidimensional arrays, called tensor time series, preserving the inherent multidimensional structure. We present a factor model approach, in a form similar to tensor CP decomposition, to the analysis of high-dimensional dynamic tensor time series. As the loading vectors are uniquely defined but not necessarily orthogonal, it is significantly different from the existing tensor factor models based on Tucker-type tensor decomposition. The model structure allows for a set of uncorrelated one-dimensional latent dynamic factor processes, making it much more convenient to study the underlying dynamics of the time series. A new high order projection estimator is proposed for such a factor model, utilizing the special structure and the idea of the higher-order orthogonal iteration procedures commonly used in the Tucker-type tensor factor model and general tensor CP decomposition procedures. Theoretical investigation provides statistical error bounds for the proposed methods, which shows the significant advantage of utilizing the special model structure.

L_2 Inference of High Dimensional Change Points Detection

Weining Wang
University of York
TBA

Online Change Point Detection in High-Dimensional Factor Models

♦*Mengyu Xu and Mahdi Mirhosseini*
University of Central Florida
In this study, we consider monitoring high-dimensional data streams for potential changes in the means. Our main assumption is that the data can be represented by the factor models with stationary second-order structure. Accordingly, we primarily deconstruct the data into the fixed-dimensional factors and high-dimensional idiosyncratic errors, and then monitor each component individually for changes. For monitoring the changes in the error component, we assume sparsity and monitoring with the MOSUM statistics. Consistency results of our proposed method is developed. Our numerical studies also confirm the applicability and performance of the methods.

Session 23INT16: Statistical Methods for Complex Medical Data

Integrative Factor Regression and Its Inference for Multimodal Data Analysis

Quefeng Li
UNC, Chapel Hill
TBC

Bayesian Analysis for Imbalanced Positive-Unlabelled Diagnosis Codes in Electronic Health Records

Ru Wang[1], ♦Ye Liang[2], Zhuqi Miao[3] and Tieming Liu[2]
[1]Dell Inc.
[2]Oklahoma State University
[3]State University of New York at New Paltz
With the increasing availability of electronic health records (EHR), significant progress has been made on developing predictive inference and algorithms by health data analysts and researchers. However, the EHR data are notoriously noisy due to missing and inaccurate inputs despite the information is abundant. One serious problem is that only a small portion of patients in the database has confirmatory diagnoses while many other patients remain undiagnosed because they did not comply with the recommended examinations. The phenomenon leads to a so-called positive-unlabelled situation and the labels are extremely imbalanced. In this paper, we propose a model-based approach to classify the unlabelled patients by using a Bayesian finite mixture model. We also discuss the label switching issue for the imbalanced data and propose a consensus Monte Carlo approach to address the imbalance issue and improve computational efficiency simultaneously. Simulation studies show that our proposed model-based approach outperforms existing positive unlabelled learning algorithms. The proposed method is applied on the Cerner EHR for detecting diabetic retinopathy (DR) patients using laboratory measurements. With only 3% confirmatory diagnoses in the EHR database, we estimate the actual DR prevalence to be 25% which coincides with reported findings in the medical literature.

Truncation Model Analysis for the under-Reporting Probability in Covid-19 Pandemic

♦*Wei Liang, Hongsheng Dai and Marialuisa Restaino*
The COVID-19 pandemic has affected all countries in the world and brought a major disruption in our daily lives. Estimation of the prevalence and contagiousness of COVID-19 infections may be challenging due to the under-reporting of infected cases in the early stage of the pandemic. For a better understanding of the current epidemic situation, it is crucial to take into consideration unreported infections. In this study, we propose a truncation model to estimate the under-reporting probabilities for infected cases. Hypothesis testing on the differences in truncation probabilities, that are related to the under-reporting rates, is implemented. Large sample results of the hypothesis test are presented theoretically and by means of simulation studies. We also apply the methodology to COVID-19 data in certain countries, where under-reporting rates are expected to be high.

Doubly Robust Methods for Selecting Optimal Treatment Based on Observational Data

Qian Xu, ♦Qi Zheng and Maiying Kong
University of Louisville
Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, where confounding often exists between treatment and outcome. In this article, we propose a flexible semiparametric outcome model to select the optimal treatment regime which is suitable for experimental studies as well as observational studies. The proposed model includes the control group response profile and the interaction term between treatment and contrast function. L1 regularization is incorporated to select the important variables in the contrast function and improve the accuracy in estimating the contrast function. The proposed approach is quite flexible and has a doubly robust nature, that is, the estimated contrast function is consistent if either the control group response profile or the propensity score model is correctly specified. Extensive simulation studies are carried out to examine the performance of the proposed method. A case study on selecting the optimal treatment to improve the inflammatory biomarker illustrates the application of the proposed method.

Session 23INT32: Real-World Challenges and Recent Developments of Statistics in Biosciences

Data Integration and Subsampling Techniques in Distribution Estimation for Event Times with Missing Origins

♦*Yi Xiong[1] and Joan Hu[2]*
[1]University of Manitoba; Fred Hutchinson Cancer Center
[2]Simon Fraser University
Time-to-event data with missing origins often arises when the occurrence of the event is silent. For example, records of wildfires can only be collected after a fire has been reported and thus the exact time when the fire starts is unknown. These temporal offsets on records of reported fires can reduce the accuracy of real-time fire prediction and monitoring. Xiong et al. (2021) proposed an approach that synthesizes auxiliary longitudinal measures to aid the inference on the unobserved time origins via the first-hitting-time model. In addition to using longitudinal measures reported after the event, one can also

consider using auxiliary information prior to the occurrence of the event. To explore the duration distribution of lightning-caused fires, we propose to use the preceding records of lightning strikes to aid inference of the ignition time and start time of a fire. We first integrate the lightning strikes data with the fire data via Kernel smoothing and then provide a distribution estimator for a fire's ignition time. By viewing a fire's start time subject to be censored within the interval of the ignition time and the report time, we further adjust Turnbull estimator with interval-censored data to estimate the distribution of missing origin. Driven by the large volume of lightning strikes data, we also adapt the proposed estimation procedures to sub-samples of the lightning strikes data. The proposed approach potentially has many applications. This research is a joint work with Professor Joan Hu in Simon Fraser University.

A Latent Variable Cace Model for Multidimensional Endpoints and Treatment Noncompliance with Application to a Longitudinal Trial of Arthritis Health Journal

♦*Lulu Guo*[1], *Joan Hu*[1], *Yi Qian*[2], *Diane Lacaille*[2] and *Hui Xie*[1]
[1]Simon Fraser University, Canada
[2]University of British Columbia, Canada

Randomized clinical trials (RCTs) are the preferred study design for assessing the causal effects of medical interventions on healthcare policymaking. Real-world RCTs evaluating multifaceted interventions often employ multiple study endpoints to measure treatment success on a small set of underlying constructs. We propose a latent variable model with principal strata of latent compliance types for parsimonious estimation of intervention effects in RCTs with multidimensional longitudinal outcomes and treatment noncompliance. Within each compliance type, a factor regression model is used to relate observed multiple endpoints to latent constructs, which are then modelled by hierarchical mixed-effects regression models. Under this model, high dimensional outcomes are reduced to low dimensional latent factors. This dimension reduction leads to a more parsimonious and efficient test of overall complier average causal effects (CACE) on multiple endpoints, mitigating the potential multiple testing issues associated with multiple endpoints. Furthermore, the inference based on factors can be more interpretable and scientifically relevant. We evaluate the performance of the proposed model using simulation studies, which shows study power can be increased substantially compared with estimating CACE for each endpoint separately. The proposed approach is illustrated by evaluating the treatment efficacy of the Arthritis Health Journal online tool. We evaluated the treatment efficacy of Arthritis Health Journal under one latent variable and two latent variables separately. Significant and beneficial treatment effects on latent variables are detected in both two situations.

A Step-Wise Multiple Testing for Linear Regression Models with Application to the Study of Resting Energy Expenditure

Junyi Zhang[1], *Zimian Wang*[2], ♦*Zhezhen Jin*[3] and *Zhiliang Ying*[4]
[1]Paul H. Chook Department of Information Systems and Statistics, Baruch College, 55 Lexington Avenue, New York, NY 10010, USA
[2]Obesity Research Center, Columbia University, New York, NY 10025, USA
[3]Department of Biostatistics, Columbia University, New York, NY 10032, USA
[4]Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

Motivated by the mechanistic model of the resting energy expenditure, we present a new multiple hypothesis testing approach to evaluate organ/tissue-specific resting metabolic rates. The approach is based on generalized marginal regression estimates for a subset of coefficients along with a stepwise multiple testing procedure with a minimization–maximization of the normalized estimates, which offers a valid way to address challenges in multiple hypothesis testing on regression coefficients in linear regression analysis especially when covariates are highly correlated. The approach also yields estimates that are conditionally unbiased and controls a family-wise error rate in the strong sense.

Data Masking with Random Orthogonal Matrix

Samuel Wu

In this talk we present application of random orthogonal matrix for privacy-preserved data collection, sharing and analysis. The new technologies can ensure that neither investigators nor analysts see the actual data, but standard statistical analysis can still be performed with almost the same results for masked data as for the original data. In addition we discuss techniques developed for data quality assurance, missing data imputation, elimination of trusted third party with a collusion resistance peer-to-peer masking. Furthermore, we show that posterior distribution of the original data given the masked data is a projection of the prior distribution onto a restricted domain, and present differential privacy properties of the new procedures.

Session 23INT38: Some Modern Issues of Statistical Learning

From Genetic Correlation Analysis to Cross-Ancestry Genetic Prediction

Lin Hou
Tsinghua University
TBC

An Integrated Deep Learning Framework for the Interpretation of Untargeted Metabolomics Data

Leqi Tian and ♦*Tianwei Yu*

Untargeted metabolomics is gaining widespread applications. The key aspects of the data analysis include modeling complex activities of the metabolic network, selecting metabolites associated with clinical outcome, and finding critical metabolic pathways to reveal biological mechanisms. One of the key roadblocks in data analysis is not well-addressed, which is the problem of matching uncertainty between data features and known metabolites. Given the limitations of the experimental technology, the identities of data features can not be directly revealed in the data. The predominant approach for mapping features to metabolites is to match the mass-to-charge ratio (m/z) of data features to those derived from theoretical values of known metabolites. The relationship between features and metabolites is not one-to-one since some metabolites share molecular composition, and various adduct ions can be derived from the same metabolite. This matching uncertainty causes unreliable metabolite selection and functional analysis results. Here we introduce an integrated deep learning framework for

metabolomics data that takes matching uncertainty into consideration. The model is devised with a gradual sparsification neural network based on the known metabolic network and the annotation relationship between features and metabolites. This architecture characterizes metabolomics data and reflects the modular structure of biological system. Three goals can be achieved simultaneously without requiring much complex inference and additional assumptions: (1) evaluate metabolite importance, (2) infer feature-metabolite matching likelihood, and (3) select disease sub-networks. When applied to a COVID metabolomics dataset and an aging mouse brain dataset, our method found metabolic sub-networks that were easily interpretable.

Robust Statistical Learning on Heavy Tailed Data

♦Lihu Xu[1], Fang Xiao[2], Qiuran Yao[1] and Huiming Zhang[1]
[1]University of Macau
[2]Peking University

There has been a surge of interest in developing robust estimators for models with heavy- tailed and bounded variance data in statistics and machine learning, while few works impose unbounded variance. This paper proposes two type of robust estimators, the ridge log- truncated M-estimator and the elastic net log-truncated M-estimator. The first estimator is applied to convex regressions such as quantile regression and generalized linear models, while the other one is applied to high dimensional non-convex learning problems such as regressions via deep neural networks. Simulations and real data analysis demonstrate the robustness of log-truncated estimations over standard estimations.

Time-Series Forecasting of Mortality Rates using Transformer

Chaojie Wang
Jiangsu University

Predicting mortality rates is a crucial issue in life insurance pricing and demographic statistics. Traditional approaches, such as the Lee-Carter model and its variants, predict the trends of mortality rates using factor models, which explain the variations of mortality rates from the perspective of ages, gender, regions, and other factors. Recently, deep learning techniques have achieved great success in various tasks and shown strong potential for time-series forecasting. In this paper, we propose a modified Transformer architecture for predicting mortality rates in major countries around the world. Through the multi-head attention mechanism and positional encoding, the proposed Transformer model extracts key features effectively and thus achieves better performance in time-series forecasting. By using empirical data from the Human Mortality Database, we demonstrate that our Transformer model has higher prediction accuracy of mortality rates than the Lee-Carter model and other classic neural networks. Our model provides a powerful forecasting tool for insurance companies and policy makers.

Session 23INT36: New Challenges in Modelling High-Dimensional and Complex Data

Multi-Kink Quantile Regression for Longitudinal Data with Application to Progesterone Data Analysis

Chuang Wan[1], ♦Wei Zhong[1], Wenyang Zhang[2] and Changliang Zou[3]

[1]Xiamen University
[2]The University of York
[3]Nankai University

Motivated by investigating the relationship between progesterone and the days in a menstrual cycle in a longitudinal study, we propose a multi-kink quantile regression model for longitudinal data analysis. It relaxes the linearity condition and assumes different regression forms in different regions of the domain of the threshold covariate. In this paper, we first propose a multi-kink quantile regression for longitudinal data. Two estimation procedures are proposed to estimate the regression coefficients and the kink points locations: one is a computationally efficient profile estimator under the working independence framework while the other one considers the within-subject correlations by using the unbiased generalized estimation equation approach. The selection consistency of the number of kink points and the asymptotic normality of two proposed estimators are established. Secondly, we construct a rank score test based on partial subgradients for the existence of kink effect in longitudinal studies. Both the null distribution and the local alternative distribution of the test statistic have been derived. Simulation studies show that the proposed methods have excellent finite sample performance. In the application to the longitudinal progesterone data, we identify two kink points in the progesterone curves over different quantiles and observe that the progesterone level remains stable before the day of ovulation, then increases quickly in five to six days after ovulation and then changes to stable again or drops slightly.

Factor Modeling for High-Dimensional Time Series with Single-Index Factor Loadings

Tao Huang
Shanghai University of Finance and Economics

Motivated by recent studies of massive high-dimensional time series data, such as COVID-19 datasets, this talk proposes a novel factor model that decomposes the dynamic behavior of high-dimensional time series data into a few low-dimensional time series factors. The factor loadings are assumed to be unknown functions of covariates, and to have a single-index form that is sufficiently flexible to avoid the problems associated with high dimensionality. An efficient iterative estimation procedure is proposed to estimate the factors and factor loadings by extending a damped Newton method. The asymptotic properties of the resulting estimators, including convergence rates and asymptotic distributions, are established. The finite sample performance of the proposed model and method is investigated using simulation studies. We also apply the proposed model and method to the analysis of the global COVID-19 dataset and an air pollutant dataset in the United States.

Functional Data Analysis for Earth Observation

Julian Austin[1], Robin Henderson[1] and ♦Jian Qing Shi[2]
[1]Newcastle University
[2]Southern University of Science and Technology
TBC

Session 23INT47: Recent Advances in Reliability

Assessing Cyber Risks of Networked Systems Based on l-Hop Propagation

Gaofeng Da
Nanjing University of Aeronautics and Astronautics

Cyber risks are the most common risks encountered by a modern networked system. Among various characteristics, propagation should be one of the most important characteristics of cyber risks. On the assumption of risk propagation, the assessment of cyber risks, namely, the size of compromised nodes after certain types of incidents, has been studied for a long time in the literature. However, most of them only focus on the limiting behaviour or approximation of size of compromised nodes of large-scale networks and ignore some important elements of risk management such as the heterogeneity, dependence and so on, which may cause a limited application. In this talk, we focus on the accurate distribution and dependence of sizes of compromised nodes in an arbitrary network based on the proposed L-hop propagation model: (1) we provide an exact algorithm for the multivariate distribution of sizes of compromised nodes with different types as well as explicit formulas; (2) we discuss the dependence of the sizes of compromised nodes with different types and we prove that it always possesses the positive association dependence among the sizes; (3) we provide some useful bounds for mean, variance and covariance of the sizes which may be used to roughly evaluate the cyber risks for a large network. Some results and useful insights for certain specific network topologies.

Non-Parametric Inference Based on Reliability Life-Test of Non-Identical Coherent Systems

Xiaojun Zhu
Xi'an Jiaotong-Liverpool University
TBC

Dependent Stress–strength Reliability of Multi-State System by Copulas using Improved Generalized Survival Signature

Xuchao Bai[1] and ♦*Mu He[2]*
[1]Xidian University
[2]Xi'an Jiaotong-Liverpool University
Multi-state systems are widely used in engineering field, which usually contain complex internal structures and variable external environmental stresses. Discussing the reliability of such system has theoretical significance and application values. In this paper, a new stress–strength model for multi-state system consisting of multi-type multi-state components is introduced with copulas. Each type of components has two dependent strengths, which is suffered from two dependent stresses in working environment, while the strengths and stresses are independent. We discuss inferential procedures for stress–strength reliability of the multi-state system using the proposed improved generalized survival signature in the case of all types of components exposed to common pair of dependent stresses.

Session 23INT15: Advances in Health and Lifetime Data Science

A Tree-Based Bayesian Accelerated Failure Time Cure Model for Estimating Heterogeneous Treatment Effect

Rongqian Sun and ♦*Xinyuan Song*
The Chinese University of Hong Kong
Estimating heterogeneous treatment effects has drawn increasing attention in medical studies, considering that patients with divergent features can undergo a different progression of disease even with identical treatment. Such heterogeneity can co-occur with a cured fraction for biomedical studies with a time-to-event outcome and further complicates the quantification of treatment effects. This study considers a joint framework of Bayesian causal forest and accelerated failure time cure model to capture the cured proportion and treatment effect heterogeneity through three separate Bayesian additive regression trees. Under the potential outcomes framework, conditional and sample average treatment effects within the uncured subgroup are derived on the scale of log survival time subject to right-censoring, and treatment effects on the scale of survival probability are derived for each individual. Bayesian backfitting Markov chain Monte Carlo algorithm with the Gibbs sampler is conducted to estimate the causal effects. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to a breast cancer dataset extracted from the SEER database to demonstrate its usage in detecting heterogeneous treatment effects and cured subgroups. Combined with popular mitigation strategies, the proposed method can also alleviate confounding induced by immortal time bias.

Optimizing Svm Parameters for High-Dimensional Class-Imbalanced Data Sets

Chen-An Tsai
National Taiwan University
For medical data mining, the development of a class prediction model has been widely used to deal with various kinds of data classification problems. Classification models especially for high dimensional gene expression datasets have attracted many researchers in order to identify marker genes for distinguishing any type of cancer cells from their corresponding normal cells. However, skewed class distributions often occur in the medical datasets in which at least one of the classes has a relatively small number of observations. A classifier induced by such an imbalanced dataset typically has a high accuracy for the majority class and poor prediction for the minority class. In this study, we focus on an SVM classifier with a Gaussian radial basis kernel for a binary classification problem. In order to take advantage of an SVM and to achieve the best generalization ability for improving the classification performance, we will address two important problems: the class imbalance and parameter selection during SVM parameter optimization. First of all, we proposed a novel adjustment method called b-SVM, for adjusting the cutoff threshold of the SVM. Second, we proposed a fast and simple approach, called the Min-max gamma selection, to optimize the model parameters of SVMs without carrying out an extensive k-fold cross validation. An extensive comparison with a standard SVM and well-known existing methods are carried out to evaluate the performance of our proposed algorithms using simulated and real datasets. The experimental results show that our proposed algorithms outperform the over-sampling techniques and existing SVM-based solutions. This study also shows that the proposed Min-max gamma selection is at least 10 times faster than the cross-validation selection based on the average running time on six real datasets.

Synthesizing Auxiliary Information in Analyzing Survival Data with Population Heterogeneity

♦*Yu-Jen Cheng[1], Yen-Chun Liu[2], Chang-Yu Tsai[1] and Chiung-Yu Huang[3]*
[1]National Tsing Hua University
[2]Duke University
[3]University of California at San Francisco

Leveraging information in aggregate data from external sources to improve estimation efficiency and prediction accuracy with smaller-scale studies has drawn a great deal of attention in recent years. Yet, conventional methods often either ignore uncertainty in the external information or fail to account for the heterogeneity between internal and external studies. This article proposes an empirical likelihood-based framework to improve the estimation of the semiparametric transformation models by incorporating information about the t-year subgroup survival probability from external sources. The proposed estimation procedure incorporates an additional likelihood component to account for uncertainty in the external information and employs a density ratio model to characterize population heterogeneity. We establish the consistency and asymptotic normality of the proposed estimator and show that it is more efficient than the conventional pseudopartial likelihood estimator without combining information. Simulation studies show that the proposed estimator yields little bias and outperforms the conventional approach even in the presence of information uncertainty and heterogeneity. The proposed methodologies are illustrated with an analysis of a pancreatic cancer study.

On a Surrogate Measure for Time-Varying Biomarkers in Randomized Clinical Trials

Ying Qing Chen
Stanford University

Clinical trials with rare or distant outcomes are usually designed to be large in size and long term. The resource-demand and time-consuming characteristics limit the feasibility and efficiency of the studies. There are motivations to replace rare or distal clinical endpoints by reliable surrogate markers, which could be earlier and easier to collect. However, statistical challenges still exist to evaluate and rank potential surrogate markers. In this paper, we define a generalized proportion of treatment effect for survival settings. The measure's definition and estimation do not rely on any model assumption. It is equipped with a consistent and asymptotically normal non-parametric estimator. Under proper conditions, the measure reflects the proportion of average treatment effect mediated by the surrogate marker among the group that would survive to mark the measurement time under both intervention and control arms.

Deep Convolutional Neural Networks for Multiclass Classification of Three Dimensional Brain Images

♦*Guan-Hua Huang, Chih-Hsuan Lin and Yu-Ren Cai*
National Yang Ming Chiao Tung University

Parkinson's disease (PD) is a degenerative disorder of the central nervous system, and it is currently diagnosed by functional medical imaging such as positron emission tomography (PET) or single photon emission computed tomography (SPECT). In this study, we use the SPECT dataset with the sample size being 634. The main objective is to develop a valid model to yield accurate prediction of PD illness stages, which is a multiclass classification task. We use the whole 3D brain imaging as the input and the overall process is fully automated. First, we treat the slices as a sequence of 2D images and put them into 2D models pretrained on ImageNet sequentially. We take the average of the outputs to yield the final predicted stage. The 3D models pretrained on Kinetics400 are also applied to our data since the SPECT imaging essentially has 3 dimensions. Age and gender are also part of the inputs of models instead of using only imaging data. Finally, we try replacing the MLP with SVM or

Random Forest classifier to increase the extent of nonlinearity. The results show that the combination of SVM/Random Forest classifier, age, and gender can help improve the accuracy and Fscore.

Session 23INT5: Recent Progresses on Change-Point Analysis

Online Kernel Cusum for Change-Point Detection

Song Wei and ♦*Yao Xie*
Georgia Institute of Technology

We propose an efficient online kernel Cumulative Sum (CUSUM) method for change-point detection that utilizes the maximum over a set of kernel statistics to account for the unknown change-point location. Our approach exhibits increased sensitivity to small changes compared to existing methods, such as the Scan-B statistic, which corresponds to a non-parametric Shewhart chart-type procedure. We provide accurate analytic approximations for two key performance metrics: the Average Run Length (ARL) and Expected Detection Delay (EDD), which enable us to establish an optimal window length on the order of the logarithm of ARL to ensure minimal power loss relative to an oracle procedure with infinite memory. Such a finding parallels the classic result for window-limited Generalized Likelihood Ratio (GLR) procedure in parametric change-point detection literature. Moreover, we introduce a recursive calculation procedure for detection statistics to ensure constant computational and memory complexity, which is essential for online procedures. Through extensive experiments on simulated data and a real-world human activity dataset, we demonstrate the competitive performance of our method and validate our theoretical results.

Detecting Multiple Anomaly Regions on Spatial Grid

♦*Chao Zheng and Baiyu Wang*
University of Southampton

There has been a growing interest in multiple change points detection problems recently, whilst focuses are mostly on changes taking place on the time index. In this project, we investigate the changes-in-mean model on a two-dimensional spatial lattice, that is, to detect the number and locations of anomaly regions from the baseline region. In addition to the usual minimisation over cost function with a penalisation related to the number of change points, we also introduce a new penalty on the diameter of the anomaly regions in the detection criterion, which limits each estimated anomaly region being too scattered. We show that the estimated number and locations of changepoints are both consistent, and characterise the error (localisation rate) of anomaly region detection based on the signal strength. We propose a numerical algorithm to solve the detection criterion and carry out Monte Carlo and real data experiments to examine the performance of our methodology.

Likelihood Score Method for Change-Points Estimation on Large-Scale Data Streams

♦*Shouri Hu[1], Jingyan Huang[2], Hao Chen[3] and Hock Peng Chan[2]*
[1]University of Electronic Science and Technology of China
[2]National University of Singapore
[3]University of California at Davis

We consider here the identification of change-points on large-scale data streams. The objective is to find the most efficient

way of combining information across data stream so that detection is possible under the smallest detectable change magnitude. The challenge comes from the sparsity of change-points when only a small fraction of data streams undergo change at any point in time. The most successful approach to the sparsity issue so far has been the application of hard thresholding such that only local scores from data streams exhibiting significant changes are considered and added. However the identification of an optimal threshold is a difficult one. In particular it is unlikely that the same threshold is optimal for different levels of sparsity. We propose here a sparse likelihood score for identifying a sparse signal. The score is a likelihood ratio for testing between the null hypothesis of no change against an alternative hypothesis in which the change-points or signals are barely detectable. By the Neyman-Pearson Lemma this score has maximum detection power at the given alternative. The outcome is that we have a scoring of data streams that is successful in detecting at the boundary of the detectable region of signals and change-points. The likelihood score can be seen as a soft thresholding approach to sparse signal and change-point detection in which local scores that indicate small changes are down-weighted much more than local scores indicating large changes. We are able to show sharp optimality of the sparsity likelihood score in the sense of achieving successful detection at the minimum detectable order of change magnitude as well as the best constant with respect this order of change.

Optimal Difference-Based Variance Estimation in Change Point Analysis and Trend Inference

Kin Wai Chan

The Chinese University of Hong Kong

Variance estimation is important for statistical inference. It becomes nontrivial when observations are masked by serial dependence structures and time-varying mean structures. Existing methods either ignore or sub-optimally handle these nuisance structures. This paper develops a general framework for the estimation of the long-run variance for time series with nonconstant means. The building blocks are difference statistics. The proposed class of estimators is general enough to cover many existing estimators. Necessary and sufficient conditions for consistency are investigated. The first asymptotically optimal estimator is derived. Our proposed estimator is theoretically proven to be invariant to arbitrary mean structures, which may include trends and a possibly divergent number of discontinuities.

Session 23INT65: Statistical Methods and Applications in High Dimensional Biological Data

Multi-Task Prediction Model for Time to Event Data

Shuai You[1], Xiaowen Cao[1], Grace Yi[2], ♦Xuekui Zhang[1] and Li Xing[3]
[1]University of Victoria
[2]University of Western Ontario
[3]University of Saskatewan
TBC

Clustering Single-Cell Rna Sequencing Data using a Mixture Model-Based Deep Learning Algorithm

Leann Lac[1], Eric Lin[2], Boyuan Liu[2], Daryl Fung[1], Carson Leung[1] and ♦Pingzhao Hu[3]
[1]University of Manitoba

[2]University of Toronto

[3]Western University

Traditional statistical methods have limitations in clustering high dimensional single-cell RNA sequencing (scRNA-seq) data which contain many biological and technical zeros. In this study, we propose a hybrid model which combines an advanced statistical model with a deep learning approach as an extendable end-to-end framework to improve the cell clustering performance on scRNA-seq data. The proposed model feeds an adjacency matrix and a gene feature matrix into a deep neural network to generate embedding. A statistical model is later applied to the embedding for cell clustering. The entire algorithm is optimized simultaneously by a designed loss function. We then apply to three labeled and three simulated scRNA-seq datasets. We consider adjusted Rand index, normalized mutual information, and Silhouette coefficient as performance metrics to evaluate the clustering performance of proposed method in comparison to four selected baseline models. The results show that the proposed method outperforms the baseline methods and has great stability in cell clustering. By successfully incorporating a statistical model into deep learning algorithm on scRNA-seq data as an end-to-end framework, the accuracy of cell clustering on scRNA-seq data is significantly improved and this improvement brings significant impact on health research.

Group Effects in Linear Models

Min Tsao

University of Victoria, Canada

In a linear regression model, parameters of predictor variables represent the individual effects of the variables. Estimating these parameters has always been an essential part of a regression analysis. When there are strongly correlated variables in the model, they generate multicollinearity which makes the least squares estimates of their parameters unreliable. In this case, a commonly used solution is to abandon the least squares regression in favor of the ridge regression or principal component regression, but these alternative methods of regression are complicated in implementation and interpretation. The sampling distributions of their estimators are also usually unavailable. In this talk, I argue that instead of abandoning the least squares regression, we should abandon the attempt to estimate parameters of the strongly correlated variables because (i) these parameters are not meaningful and (ii) they cannot be accurately estimated regardless of the method of regression used unless the sample size is very large. How, then, do we analyze such variables? I propose that we analyze their collective impact on the response variable. To this end, I introduce group effects, a class of linear combinations of their parameters, to represent their collective impact, and show that there are group effects that can be accurately estimated by their minimum-variance unbiased linear estimators. These group effects also have good interpretations and characterize the region of the predictor variable space where the least squares estimated model makes accurate predictions. They provide a means to study strongly correlated variables through the simple least squares regression, which is useful for analyzing observational data from social science, environmental science, and medical research where strongly correlated variables are common.

Session 23INT17: New Statistical Methods for Analyzing Complex Survival Data

Longitudinal Variable Selection under Joint Modeling of Longitudinal and Survival Data

♦*Fangfang Wang*[1], *Lu Lin*[2], *Lei Liu*[3], *Hongmei Jiang*[4] *and Lihui Zhao*[4]
[1]Yancheng Institute of Technology
[2]Shandong University
[3]Washington University in St. Louis
[4]Northwestern University
TBC

Scalable Estimation for High Velocity Survival Data

♦*Ying Sheng*[1], *Yifei Sun*[2], *Charles E. Mcculloch*[3] *and Chiung-Yu Huang*[3]
[1]Chinese Academy of Sciences
[2]Columbia University
[3]University of California at San Francisco
With the rapidly increasing availability of large-scale streaming data, there has been a growing interest in developing methods that allow the processing of the data in batches without requiring storage of the full dataset. In this paper, we propose a hybrid likelihood approach for scalable estimation of the Cox model using individual-level data in the current data batch and summary statistics calculated from historical data. We show that the proposed scalable estimator is asymptotically as efficient as the maximum likelihood estimator calculated using the entire dataset with low data storage requirements and low loading and computation time. A challenge in analyzing survival data batches that is not accommodated in extant methods is that new covariates may become available midway through data collection. To accommodate addition of covariates, we develop a hybrid empirical likelihood approach to incorporate the historical covariate effects evaluated in a reduced Cox model. The extended scalable estimator is asymptotically more efficient than the maximum likelihood estimator obtained using only the data batches that include the additional covariates. The proposed approaches are evaluated by numerical simulations and illustrated with an analysis of Surveillance, Epidemiology, and End Results (SEER) breast cancer data. This is a joint work with Yifei Sun, Charles E. McCulloch, and Chiung-Yu Huang

A Semiparametric Cox-Aalen Transformation Model with Censored Data

Xi Ning[1], *Yinghao Pan*[2], ♦*Yanqing Sun*[2] *and Peter Gilbert*[3]
[1]University of North Carolina at Charlotte
[2]University of North Carolina at Charlotte, USA
[3]Fred Hutchinson Cancer Center and University of Washington
We propose a broad class of so-called Cox-Aalen transformation models that incorporate both multiplicative and additive covariate effects on the baseline hazard function within a transformation. The proposed models provide a highly flexible and versatile class of semiparametric models that include the transformation models and the Cox-Aalen model as special cases. Specifically, it extends the transformation models by allowing potentially time-dependent covariates to work additively on the baseline hazard and extends the Cox-Aalen model through a predetermined transformation function. We propose an estimating equation approach and devise an Expectation-Solving (ES) algorithm that involves fast and robust calculations. The resulting estimator is shown to be consistent and asymptotically normal via modern empirical process techniques. The ES algorithm yields a computationally simple method for estimating the variance of both parametric and nonparametric estimators. Finally, we demonstrate the performance of our procedures through extensive simulation studies and applications in two randomized, placebo-controlled HIV prevention efficacy trials. The data example shows the utility of the proposed Cox-Aalen transformation models in enhancing statistical power for discovering covariate effects.

Conditional Quasi-Likelihood Inference for Mean Residual Life Regression with Clustered Failure Time Data

Rui Huang and ♦*Liming Xiang*
Nanyang Technological University, Singapore
Cox frailty models have been widely used to investigate the correlation of failure time data within clusters. In this work, we propose to analyze clustered failure time data under a frailty proportional mean residual life regression model using a novel conditional quasi-likelihood inference procedure. The proposed method utilizes the stochastic process and the inverse probability of censoring weighting to form quasi-scores for regression parameters. Conditional inference based on a penalized quasi-likelihood is developed to address within-cluster correlation without specifying the distribution for frailty, bringing the method closer to what suffices for real-world applications. We establish the asymptotic properties of the proposed estimator, and evaluate the performance of this new proposal via simulation studies and a real data example.

Session 23INT46: Recent Developments in Statistical Machine Learning

Regularized Greedy Gradient q-Learning with Mobile Health Applications.

Min Qian
Columbia University
Recent advance in health and technology has made mobile apps a viable approach to delivering behavioral interventions in areas including physical activity encouragement, smoking cessation, substance abuse prevention, and mental health management. Due to the chronic nature of most of the disorders and heterogeneity among mobile users, delivery of the interventions needs to be sequential and tailored to individual needs. We operationalize the sequential decision making via a policy that takes a mobile user's past usage pattern and health status as input and outputs an app/intervention recommendation with the goal of optimizing the cumulative rewards of interest in an indefinite horizon setting. we propose a regularized greedy gradient Q-learning (RGGQ) method to tackle this estimation problem. The optimal policy is estimated via an algorithm which synthesizes the PGM and the GGQ algorithms, and its asymptotic properties are established.

Differential Privacy in Personalized Pricing with Nonparametric Demand Models

Xi Chen[1], *Sentao Miao*[2] *and* ♦*Yining Wang*[3]
[1]New York University
[2]McGill University
[3]University of Texas at Dallas

In the recent decades, the advance of information technology and abundant personal data facilitate the application of algorithmic personalized pricing. However, this leads to the growing concern of potential violation of privacy due to adversarial attack. To address the privacy issue, this paper studies a dynamic personalized pricing problem with *unknown* nonparametric demand models under data privacy protection. Two concepts of data privacy, which have been widely applied in practices, are introduced: *central differential privacy (CDP)* and *local differential privacy (LDP)*, which is proved to be stronger than CDP in many cases. We develop two algorithms which make pricing decisions and learn the unknown demand on the fly, while satisfying the CDP and LDP gurantees respectively. In particular, for the algorithm with CDP guarantee, the regret is proved to be at most $\tilde{O}(T^{(d+2)/(d+4)} + \varepsilon^{-1}T^{d/(d+4)})$. Here, the parameter T denotes the length of the time horizon, d is the dimension of the personalized information vector, and the key parameter $\varepsilon > 0$ measures the strength of privacy (smaller ε indicates a stronger privacy protection). On the other hand, for the algorithm with LDP guarantee, its regret is proved to be at most $\tilde{O}(\varepsilon^{-2/(d+2)}T^{(d+1)/(d+2)})$, which is near-optimal as we prove a lower bound of $\Omega(\varepsilon^{-2/(d+2)}T^{(d+1)/(d+2)})$ for any algorithm with LDP guarantee.

Learning Linear Non-Gaussian Dag with Diverging Number of Nodes
Junhui Wang
Chinese University of Hong Kong
An acyclic model, often depicted as a directed acyclic graph (DAG), has been widely employed to represent directional causal relations among collected nodes. In this talk, we present an efficient method to learn linear non-Gaussian DAG in high dimensional cases, where the noises can be of any continuous non-Gaussian distribution. The proposed method leverages the concept of topological layer to facilitate the DAG learning, and its theoretical justification in terms of exact DAG recovery is also established under mild conditions. Particularly, we show that the topological layers can be exactly reconstructed in a bottom-up fashion, and the parent-child relations among nodes can also be consistently established. The established asymptotic DAG recovery is in sharp contrast to that of many existing learning methods assuming parental faithfulness or ordered noise variances. The advantage of the proposed method is also supported by the numerical comparison against some popular competitors in various simulated examples as well as a real application on the global spread of COVID-19.

Online Estimation with Dependent Samples and Robust Policy Evaluation in Reinforcement Learning
Xi Chen[1], Weidong Liu[2], Jiyuan Tu[2] and ♦Yichen Zhang[3]
[1]New York University
[2]Shanghai Jiao Tong University
[3]Purdue University
We propose a robust policy evaluation algorithm in reinforcement learning, to feature outlier contamination and heavy-tailed reward distributions. We further develop a fully-online method to conduct statistical inference. Our method converges faster to the minimum asymptotic variance than the classical temporal difference (TD) learning and avoids the selection of the step sizes. Numerical experiments are provided on the effectiveness of the proposed algorithm in real-world reinforcement learning experiments.

Session 23INTKT2: Keynote Talk 2: Qiman Shao

Perspective of Self-Normalized Limit Theory
Qi-Man Shao
Southern University of Science and Technology
Limit theory plays an important role in probability and statistics. Classi- cal limit theorems such as the law of large numbers , the central limit theorem and the Cra´mer moderate deviation theorem, under deterministic standard- ization, have been well developed and understood. However, standardized coefficients in applications are more often random, or self-normalized. In this talk, we shall review recent developments of limit theory for self-normalized processes as well as applications to statistical inference.

Session 23INT100: Network Modeling and Applications

Statistics About Statisticians
Jiashun Jin
TBC

Individual-Centered Partial Information in Social Networks
Xiao Han[1], Rachel Wang[2] and ♦Xin Tong[3]
[1]University of Science and Technology of China
[2]University of Sydney
[3]University of Southern California
In statistical network analysis, we often assume either the full network is available or multiple subgraphs can be sampled to estimate various global properties of the network. However, in a real social network, people frequently make decisions based on their local view of the network alone. Here, we consider a partial information framework that characterizes the local network centered at a given individual by path length L and gives rise to a partial adjacency matrix. Under L=2, we focus on the problem of (global) community detection using the popular stochastic block model (SBM) and its degree-corrected variant (DCSBM). We derive general properties of the eigenvalues and eigenvectors from the signal term of the partial adjacency matrix and propose new spectral-based community detection algorithms that achieve consistency under appropriate conditions. Our analysis also allows us to propose a new centrality measure that assesses the importance of an individual's partial information in determining global community structure. Using simulated and real networks, we demonstrate the performance of our algorithms and compare our centrality measure with other popular alternatives to show it captures unique nodal information. Our results illustrate that the partial information framework enables us to compare the viewpoints of different individuals regarding the global structure.

Group Network Hawkes Process
Guanhua Fang[1], Ganggang Xu[2], Haochen Xu[1], ♦Xuening Zhu[1] and Yongtao Guan[2]
[1]Fudan University
[2]University of Miami
In this work, we study the event occurrences of individuals interacting in a network. To characterize the dynamic interactions among the individuals, we propose a group network Hawkes process (GNHP) model whose network structure is observed and fxed. In particular, we introduce a latent group structure

among individuals to account for the heterogeneous user-specifc characteristics. A maximum likelihood approach is proposed to simultaneously cluster individuals in the network and estimate model parameters. A fast EM algorithm is subsequently developed by utilizing the branching representation of the proposed GNHP model. Theoretical properties of the resulting estimators of group memberships and model parameters are investigated under both settings when the number of latent groups G is over-specifed or correctly specifed. A data-driven criterion that can consistently identify the true G under mild conditions is derived. Extensive simulation studies and an application to a data set collected from Sina Weibo are used to illustrate the effectiveness of the proposed methodology.

Inference for Heteroscedastic Pca with Missing Data

♦Jianqing Fan and Yihong Gu
Princeton University

This paper studies how to construct confidence regions for principal component analysis (PCA) in high dimension, a problem that has been vastly under-explored. While computing measures of uncertainty for nonlinear / nonconvex estimators is in general difficult in high dimension, the challenge is further compounded by the prevalent presence of missing data and heteroskedastic noise. We propose a novel approach to performing valid inference on the principal subspace, on the basis of an estimator called HeteroPCA (Zhang et al., 2022). We develop non-asymptotic distributional guarantees for HeteroPCA, and demonstrate how these can be invoked to compute both confidence regions for the principal subspace and entrywise confidence intervals for the spiked covariance matrix. Our inference procedures are fully data-driven and adaptive to heteroskedastic random noise, without requiring prior knowledge about the noise levels.

Session 23INT49: Recent Developments in Deep Learning

Conditional Stochastic Interpolation: a New Approach to Conditional Sampling

♦Ding Huang, Guohao Shen, Ting Li and Jian Huang
The Hong Kong Polytechnic University

We present a novel framework, called the Conditional Stochastic Interpolation, which learns the probability flow equations or stochastic differential equations to transport between two empirically observed distributions. The proposed framework provides an unified solution to conditional generative modeling and domain transfer. The key idea is to learn the velocity function and the conditional score function basing on the conditional stochastic interpolation, which are then used to construct a Markov process for the purpose of conditional sampling. We also derive an upper bound for the excess risk of the estimation via ReLU activated neural networks. To the best of our knowledge, this is the first systematic study on stochastic interpolation with conditions. Numerical studies confirm our theoretical findings. The method is applied to study the aging of human brains, yielding interesting findings.

Robust Structure Learning and L_p-Regularization for Graph Neural Networks.

Shaogao Lv
TBC

Optimal Rates of Approximation by Shallow Reluʀ Neural Networks and Applications to Nonparametric Regression

♦Yunfei Yang[1] and Ding-Xuan Zhou[2]
[1]City University of Hong Kong
[2]University of Sydney

In the first part of this talk, we discuss the approximation capacity of some variation spaces corresponding to shallow ReLUʀ neural networks. We show that sufficiently smooth functions are contained in these spaces with finite variation norms. For functions with less smoothness, the approximation rates in terms of the variation norm are established. Using these results, we are able to prove the optimal approximation rates in terms of the number of neurons for shallow ReLUʀ neural networks. In the second part, we apply these approximation results to study convergence rates of nonparametric regression using three ReLU neural network models: shallow neural network, over-parameterized neural network, and CNN. In particular, we show that shallow neural networks can achieve the minimax optimal rates for learning Hölder functions, which complements recent results for deep neural networks. It is also proven that over-parameterized (deep or shallow) neural networks can achieve nearly optimal rates for nonparametric regression.

Deep Sufficient Representation Learning via Mutual Information

♦Siming Zheng[1], Yuanyuan Lin[1] and Jian Huang[2]
[1]The Chinese University of Hong Kong
[2]The Hong Kong Polytechnic University

We propose a mutual information-based sufficient representation (MISR) learning approach, which uses the variational formulation of mutual information and leverages the approximation power of deep neural networks. MISR learns a sufficient representation with the maximum mutual information with the response and a user-selected distribution. It can easily handle multi-dimensional continuous or categorical response variables. MISR is shown to be consistent in the sense that the conditional probability density function of the response variable given the learned representation converges to the conditional probability density function of the response variable given the predictor. Non-asymptotic error bounds for MISR are also established under suitable conditions. We evaluate MISR via extensive numerical studies with simulated and real data. The results from the numerical studies suggest that MISR outperforms several existing sufficient dimension reduction methods. We also demonstrate using the MNIST dataset that MISR works well in learning an effective representation with the full dataset but it is computationally prohibitive to use two existing nonparametric dimension reduction methods.

Error Analysis for Deep Adversarial Traning

Yuling Jiao

In order to assess the effectiveness of the adversarial estimator, we examine its ability to generalize by measuring the adversarial excess risk. Our analysis method involves examining both the generalization error and approximation error. Our findings demonstrate that the adversarial loss is robust in terms of its reliance on the level of adversarial attacks. Through a bias-variance-robust trade-off approach, we have established non-asymptotic upper bounds for the adversarial excess risk associated with Lipschitz loss functions. Furthermore, we have applied our general results to adversarial training for both classification and regression problems.

Session 23INT44: Recent Advances in Matrix and Tensor Data Analysis

Inference for Heteroskedastic Pca with Missing Data

Yuling Yan[1], ♦Yuxin Chen[2] and Jianqing Fan[1]
[1]Princeton University
[2]University of Pennsylvania

This work studies how to construct confidence regions for principal component analysis (PCA) in high dimension, a problem that has been vastly under-explored. While computing measures of uncertainty for nonlinear/nonconvex estimators is in general difficult in high dimension, the challenge is further compounded by the prevalent presence of missing data and heteroskedastic noise. We propose a suite of solutions to perform valid inference on the principal subspace based on two estimators: a vanilla SVD-based approach, and a more refined iterative scheme called HeteroPCA. We develop non-asymptotic distributional guarantees for both estimators, and demonstrate how these can be invoked to compute both confidence regions for the principal subspace and entrywise confidence intervals for the spiked covariance matrix. Particularly worth highlighting is the inference procedure built on top of HeteroPCA, which is not only valid but also statistically efficient for broader scenarios (e.g., it covers a wider range of missing rates and signal-to-noise ratios). Our solutions are fully data-driven and adaptive to heteroskedastic random noise, without requiring prior knowledge about the noise levels and noise distributions.

Tensor t Distribution and Tensor Response Regression

♦*Ning Wang[1], Qing Mai[2] and Xin Zhang[2]*
[1]Beijing Normal University
[2]Florida State University

In recent years, promising statistical modeling approaches to tensor data analysis have been rapidly developed. Traditional multivariate analysis tools, such as multivariate regression and discriminant analysis, are now generalized from modeling random vectors and matrices to higher-order random tensors (a.k.a. array-valued random objects). Equipped with tensor algebra and high-dimensional computation techniques, concise and interpretable statistical models and estimation procedures prevail in many applications. One of the biggest challenges to statistical tensor models is the non-Gaussian nature of many real-world data. Unfortunately, existing approaches are either restricted to normality or implicitly using least squares type objective functions that are computationally efficient but sensitive to data contamination. Motivated by this, we propose a simple tensor t-distribution that is, unlike existing matrix t-distributions, compatible with tensor operators and reshaping of the data. We then study the tensor response regression with tensor t-error, and develop penalized estimation and hypothesis testing under this t-modeling approach. A novel one-step estimation algorithm is developed for penalized non-convex optimization and is proven to converge to the global optimum. We study the asymptotic relative efficiency of various estimators under this model and establish the oracle properties in variable selection and near-optimal asymptotic efficiency. Extensive numerical studies show the encouraging performance of the one-step estimator.

Matrix Completion with Model-Free Weighting

Jiayi Wang[1], ♦Raymond K. W. Wong[1], Xiaojun Mao[2] and Kwun Chuen Gary Chan[3]
[1]Texas A&M University
[2]Fudan University
[3]University of Washington

In this work, we propose a novel method for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys appealing theoretical guarantees. In particular, the proposed method achieves stronger guarantee than existing work in terms of the scaling with respect to the observation probabilities, under asymptotically heterogeneous missing settings (where entry-wise observation probabilities can be of different orders). These settings can be regarded as a better theoretical model of missing patterns with highly varying probabilities. We also provide a new minimax lower bound under a class of heterogeneous settings. Numerical experiments are also provided to demonstrate the effectiveness of the proposed method.

Session 23INT96: Recent Developments in Genetics and Genomics and High Dimensional Data

Improving Polygenic Risk Prediction in Admixed Populations by Explicitly Modeling Ancestral-Specific Effects via Gaudi

Yun Li
University of North Carolina at Chapel Hill

Polygenic risk scores (PRS) have shown successes in clinics, but most PRS methods have focused only on individuals with one primary continental ancestry, thus poorly accommodating recently-admixed individuals. Here, we develop GAUDI, a novel penalized-regression-based method specifically designed for admixed individuals by explicitly modeling ancestry-specific effects and jointly estimating ancestry-shared effects. We demonstrate marked advantages of GAUDI over other methods through comprehensive simulation and real data analyses.

Machine-Learning-Based Genotype Imputation Quality Calibration

♦*Quan Sun[1], Yingxi Yang[2], Jonathan Rosen[1], Laura Raffield[1], Michael Bamshad[3], Garry Cutting[4], Michael Knowles[1], Daniel Schrider[1], Christian Fuchsberger[5] and Yun Li[1]*
[1]University of North Carolina at Chapel Hill
[2]Yale University
[3]University of Washington
[4]Johns Hopkins University
[5]EURAC Research

Whole-genome sequencing (WGS) is the gold standard for fully characterizing genetic variation but is still prohibitively expensive for large samples. To reduce costs, many studies sequence only a subset of individuals or genomic regions, and genotype imputation is used to infer genotypes for the remaining individuals or regions without sequencing data. However, not all variants can be well imputed, and the current state-of-the-art imputation quality metric, denoted as standard Rsq, is poorly calibrated for lower-frequency variants. Here, we propose MagicalRsq, a machine-learning-based method that integrates variant-level imputation and population genetics statistics, to

provide a better calibrated imputation quality metric. Leveraging WGS data from the Cystic Fibrosis Genome Project (CFGP), and whole-exome sequence data from UK BioBank (UKB), we performed comprehensive experiments to evaluate the performance of MagicalRsq compared to standard Rsq for partially sequenced studies. We found that MagicalRsq aligns better with true R2 than standard Rsq in almost every situation evaluated, for both European and African ancestry samples. For example, when applying models trained from 1,992 CFGP sequenced samples to an independent 3,103 samples with no sequencing but TOPMed imputation from array genotypes, MagicalRsq, compared to standard Rsq, achieved net gains of 1.4 million rare, 117k low-frequency, and 18k common variants, where net gains were gained numbers of correctly distinguished variants by MagicalRsq over standard Rsq. MagicalRsq can serve as an improved post-imputation quality metric and will benefit downstream analysis by better distinguishing well-imputed variants from those poorly imputed.

High-Dimensional Robust Inference for Censored Linear Models

Jiayu Huang[1] *and* ♦*Yuanshan Wu*[2]
[1]Wuhan University
[2]Zhongnan University of Economics and Law
Due to the directly statistical interpretation, censored linear regression offers a valuable complement to the Cox proportional hazards regression in survival analysis. We propose a rank-based high-dimensional inference for censored linear regression without imposing any moment condition on the model error. We develop theory of high-dimensional U-statistic, circumvent challenges stemming from the non-smoothness of loss function, and establish convergence rate of regularized estimator and asymptotic normality of the resulting de-biased estimator as well as consistency of the asymptotic variance estimation. As censoring can be viewed as a manner of trimming, it thereby strengthens the robustness of the rank-based high-dimensional inference, particularly for heavy-tailed model error or outlier in the presence of response. We evaluate the finite-sample performance of the proposed method via extensive simulation studies and demonstrate its utility by applying it to a subcohort study from The Cancer Genome Atlas (TCGA).

Statistical Method for Tct Data

Si Liu, Phil Bradley and ♦*Wei Sun*
Fred Hutchinson Cancer Center
A T cell relies on its T cell receptor (TCR) to recognize foreign antigens presented by a human leukocyte antigen (HLA), which is the human version of major histocompatibility complex (MHC). HLA is the most polymorphic locus in human genome. We explore the capacity of neural networks to predict the association between HLA and TCR, based on their amino acid sequences. We quantify the functional similarities of HLA alleles based on the predictions of TCR-HLA associations, and demonstrate the association of such similarities with survival outcome of cancer patients who received immune checkpoint blockade (ICB) treatment.

Session 23INT28: Functional and Metric Space Data

Dimension Reduction for Frechet Regression

Qi Zhang, Lingzhou Xue and ♦*Bing Li*
Pennsylvania State University

With the rapid development of data collection techniques, complex data objects that are not in the Euclidean space are frequently encountered in new statistical applications. Frechet regression model (Peterson & Müller 2019) provides a promising framework for regression analysis with metric space-valued responses. In this paper, we introduce a flexible sufficient dimension reduction (SDR) method for Fréchet regression to achieve two purposes: to mitigate the curse of dimensionality caused by high-dimensional predictors and to provide a visual inspection tool for Fréchet regression. Our approach is flexible enough to turn any existing SDR method for Euclidean (X,Y) into one for Euclidean X and metric space-valued Y. The basic idea is to first map the metric-space valued random object Y to a real-valued random variable f(Y) using a class of functions, and then perform classical SDR to the transformed data. If the class of functions is sufficiently rich, then we are guaranteed to uncover the Fréchet SDR space. We showed that such a class, which we call an ensemble, can be generated by a universal kernel. We established the consistency and asymptotic convergence rate of the proposed methods. The finite-sample performance of the proposed methods is illustrated through simulation studies for several commonly encountered metric spaces that include Wasserstein space, the space of symmetric positive definite matrices, and the sphere. We illustrated the data visualization aspect of our method by exploring the human mortality distribution data across countries and by studying the distribution of hematoma density.

A Unified Approach to Hypothesis Testing for Functional Linear Models

Yinan Lin and ♦*Zhenhua Lin*
National University of Singapore
We develop a unified approach to hypothesis testing for various types of widely used functional linear models, such as scalar-on-function, function-on-function and function-on-scalar models. In addition, the proposed test applies to models of mixed types, such as models with both functional and scalar predictors. In contrast with most existing methods that rest on the large-sample distributions of test statistics, the proposed method leverages the technique of bootstrapping max statistics and exploits the variance decay property that is an inherent feature of functional data, to improve the empirical power of tests especially when the sample size is limited and the signal is relatively weak. Theoretical guarantees on the validity and consistency of the proposed test are provided uniformly for a class of test statistics.

Geometric Eda for Random Objects

♦*Paromita Dubey*[1]*, Yaqing Chen*[2] *and Hans-Georg Müller*[3]
[1]USC
[2]Rutgers University
[3]UC Davis
In this talk I will propose new tools for the exploratory data analysis of data objects taking values in a general separable metric space. First, I will introduce depth profiles, where the depth profile of a point in the metric space refers to the distribution of the distances between and the data objects. I will describe how depth profiles can be harnessed to define transport ranks, which capture the centrality of each element in the metric space with respect to the data cloud. Next, I will discuss the properties of transport ranks and show how they can be an effective device for detecting and visualizing patterns in samples

of random objects. Together with practical illustrations I will establish the theoretical guarantees for the estimation of the depth profiles and the transport ranks for a wide class of metric spaces. I will demonstrate the efficacy of this new approach on distributional data comprising of a sample of age-at-death distributions for various countries, for compositional data through energy usage for the U.S. states and for neuroimaging network data. This talk is based on joint work with Yaqing Chen and Hans-Georg Müller.

Geometric Exploration of Random Objects Through Optimal Transport

Paromita Dubey[1], *♦Yaqing Chen*[2] *and Hans-Georg Müller*[3]
[1]University of Southern California
[2]Rutgers University
[3]University of California, Davis

We propose new tools for the geometric exploration of data objects taking values in a general separable metric space. For a random object, we first introduce the concept of depth profiles. Specifically, the depth profile of a point in a metric space is the distribution of distances between the very point and the random object. Depth profiles can be harnessed to define transport ranks based on optimal transport, which capture the centrality and outlyingness of each element in the metric space with respect to the probability measure induced by the random object. We study the properties of transport ranks and show that they provide an effective device for detecting and visualizing patterns in samples of random objects. In particular, we establish the theoretical guarantees for the estimation of the depth profiles and the transport ranks for a wide class of metric spaces, followed by practical illustrations.

Session 23INT48: Recent Developments in Statistical Genomics with Applications to COVID-19

Differential Inference for Single-Cell Rna-Sequencing Data

♦Fangda Song[1], *Kevin Yip*[2] *and Yingying Wei*[2]
[1]The Chinese University of Hong Kong, Shenzhen
[2]The Chinese University of Hong Kong

With the wide adoption of single-cell RNA-seq (scRNA-seq) technologies, scRNA-seq experiments are becoming more and more complicated with multiple treatments or biological conditions. However, despite the active research on batch effects correction, cell type clustering, and missing data imputation for scRNA-seq data, rigorous statistical methods to compare scRNA-seq experiments under different conditions are still lacking. Here, we propose a Bayesian hierarchical model, DIFferential Inference for Single-cell RNA-sequencing Data (DIFseq), to rigorously quantify the treatment effects on both cellular compositions and cell-type-specific gene expression levels for scRNA-seq data. We derive conditions for the model identifiability, which provides guidelines on the experimental design for comparative scRNA-seq studies. We implement a highly scalable Monte Carlo Expectation-Maximization algorithm to handle a large number of cells. Application of DIFseq to a pancreatic study demonstrates that considering the biological conditions of samples in the analysis substantially boosts the clustering accuracy as compared to traditional analysis pipelines for scRNA-seq data and identifies cell-type-specific and condition-specific differentially expressed genes.

Identification of Cell-Type-Specific Spatially Variable Genes Accounting for Excess Zeros

Xiangyu Luo
Renmin University of China

Motivation Spatial transcriptomic techniques can profile gene expressions while retaining the spatial information, thus offering unprecedented opportunities to explore the relationship between gene expression and spatial locations. The spatial relationship may vary across cell types, but there is a lack of statistical methods to identify cell-type-specific spatially variable (SV) genes by simultaneously modeling excess zeros and cell-type proportions. Results We develop a statistical approach CTSV to detect cell-type-specific SV genes. CTSV directly models spatial raw count data and considers zero-inflation as well as overdispersion using a zero-inflated negative binomial distribution. It then incorporates cell-type proportions and spatial effect functions in the zero-inflated negative binomial regression framework. The R package pscl is employed to fit the model. For robustness, a Cauchy combination rule is applied to integrate P-values from multiple choices of spatial effect functions. Simulation studies show that CTSV not only outperforms competing methods at the aggregated level but also achieves more power at the cell-type level. By analyzing pancreatic ductal adenocarcinoma spatial transcriptomic data, SV genes identified by CTSV reveal biological insights at the cell-type level. Availability and implementation The R package of CTSV is available at https://bioconductor.org/packages/devel/bioc/html/CTSV.html. Supplementary information Supplementary data are available at Bioinformatics online.

A Novel Penalized Inverse-Variance Weighted Estimator for Mendelian Randomization with Applications to Covid-19 Outcomes

Siqi Xu[1], *Peng Wang*[2], *Wing Kam Fung*[1] *and ♦Zhonghua Liu*[3]
[1]University of Hong Kong
[2]Huazhong University of Science and Technology
[3]Columbia University

Mendelian randomization utilizes genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome of interest even in the presence of unmeasured confounders. However, the popular inverse-variance weighted (IVW) estimator could be biased in the presence of weak IVs, a common challenge in MR studies. In this article, we develop a novel penalized inverse-variance weighted (pIVW) estimator, which adjusts the original IVW estimator to account for the weak IV issue by using a penalization approach to prevent the denominator of the pIVW estimator from being close to zero. Moreover, we adjust the variance estimation of the pIVW estimator to account for the presence of balanced horizontal pleiotropy. We show that the recently proposed debiased IVW (dIVW) estimator is a special case of our proposed pIVW estimator. We further prove that the pIVW estimator has smaller bias and variance than the dIVW estimator under some regularity conditions. We also conduct extensive simulation studies to demonstrate the performance of the proposed pIVW estimator. Furthermore, we apply the pIVW estimator to estimate the causal effects of five obesity-related exposures on three coronavirus disease 2019 (COVID-19) outcomes. Notably, we find that hypertensive disease is associated with an increased risk of hospitalized COVID-19; and peripheral vascular disease and higher body mass index are associated with increased risks of

COVID-19 infection, hospitalized COVID-19, and critically ill COVID-19.

Subtyping of Major Sars-Cov-2 Variants Reveals Different Transmission Dynamics Based on 10 Million Genomes

♦Hsin-Chou Yang[1], Jen-Hung Wang[1], Chih-Ting Yang[1], Yin-Chun Lin[1], Han-Ni Hsieh[1], Po-Wen Chen[1], Hsiao-Chi Liao[1], Chun-Houh Chen[1] and James C. Liao[2]
[1]Institute of Statistical Science, Academia Sinica
[2]Institute of Biological Chemistry, Academia Sinica

SARS-CoV-2 continues to evolve, causing waves of the pandemic. Up to May 2022, 10 million genome sequences have accumulated, which are classified into five major variants of concern. With the growing number of sequenced genomes, analysis of the big dataset has become increasingly challenging. Here we developed systematic approaches based on sets of correlated single nucleotide variations (SNVs) for comprehensive subtyping and pattern recognition of transmission dynamics. The approach outperformed single-SNV and spike-centric scans. Moreover, the derived subtypes elucidate the relationship of signature SNVs and transmission dynamics. We found that different subtypes of the same variant including Delta and Omicron exhibited distinct temporal trajectories. For example, some Delta and Omicron subtypes did not spread rapidly, while others did. We identified sets of characteristic SNVs that appeared to enhance transmission or decrease efficacy of antibodies for some subtypes. We also identified a set of SNVs that appeared to suppress transmission or increase viral sensitivity to antibodies. For the Omicron variant, the dominant type in the world, we identified the subtypes with enhanced and suppressed transmission in an analysis of eight million genomes as of March 2022 and further confirmed the findings in a later analysis of ten million genomes as of May 2022. While the "enhancer" SNVs exhibited an enriched presence on the spike protein, the "suppressor" SNVs are mainly elsewhere. Disruption of the SNV correlation largely destroyed the enhancer-suppressor phenomena. These results suggest the importance of fine subtyping of variants, and point to potential complex interactions among SNVs.

Session 23INT61: Recent Developments on Statistical Inference and Clustering

Inferring Social Influence in Dynamic Networks

Yuguo Chen
University of Illinois Urbana-Champaign

An interesting problem in social network analysis is whether individuals' behaviors or opinions spread from one to another, which is known as social influence. The degrees of influence describes how far the influence passes through individuals. Here, we explore the degrees of influence in dynamic networks. We build a longitudinal influence model to specify how people's behaviors are influenced by others in a dynamic network. In order to determine the degrees of influence, we propose a sequential hypothesis testing procedure and use generalized estimating equations to account for multiple observations of the same individual across different time points. In addition, we show that the power of our proposed test goes to one as the network size goes to infinity. We illustrate the performance of our proposed method using simulation studies and real-data analyses.

Bayesian Biclustering and Its Application in Education Data Analysis

Weining Shen
TBC

Bootstrap the Cross-Validation Estimator

Bryan Cai[1], Fabio Pellegrini[2], Menglan Pang[2], Car De Moor[2], Changyu Shen[2], Vivek Charu[1] and ♦Lu Tian[1]
[1]Stanford University
[2]Biogen

Cross-validation is a widely used technique for evaluating the performance of prediction models. It helps avoid the optimism bias in error estimate, which can be significant for models built using complex statistical learning algorithms. The cross-validation estimate, however, is a random value that depends on the observed data. It is crucial to appropriately quantify the uncertainty of the cross-validation estimate to better use the cross-validation result in practice. For instance, when comparing the performance of two prediction models using cross-validation, it's important to know if the observed difference in prediction error estimate could simply be a result of random fluctuation in the data. Although various methods have been developed for making inferences on cross-validation estimates, they often have many limitations including stringent model assumptions. In this paper, we propose a fast bootstrap method to estimate the standard error of the cross-validation estimate and construct valid confidence intervals for a population parameter measuring average model performance. Our method overcomes the computational challenge in bootstrapping the cross-validation estimate by estimating the variance component in a random effects model. It is as flexible as the cross-validation procedure itself. To demonstrate the effectiveness of the new method, we use three diverse applications, including comprehensive simulations and real data analysis examples.

Session 23INT50: Real World Evidence Research in Pharmaceutical Development

Optimal Treatment Regimes: a Review and Empirical Comparison

Zhen Li[1], ♦Jie Chen[2], Eric Laber[3], Fang Liu[4] and Richard Baumgartner[4]
[1]Amazon
[2]ECR Global
[3]Duke
[4]Merck

A treatment regime is a sequence of decision rules, one per decision point, that maps accumulated patient information to a recommended intervention. An optimal treatment regime maximizes expected cumulative utility if applied to select interventions in a population of interest. As a treatment regime seeks to improve the quality of healthcare by individualizing treatment, it can be viewed as an approach to formalizing precision medicine. Increased interest and investment in precision medicine has led to a surge of methodological research focusing on estimation and evaluation of optimal treatment regimes from observational and/or randomized studies. These methods are becoming commonplace in biomedical research, though guidance about how to choose among existing methods in practice has been somewhat limited. The purpose of this review is to describe some of the most commonly used methods for estimation

of an optimal treatment regime, and to compare these estimators in a series of simulation experiments and applications to real data. The results of these simulations along with the theoretical/methodological properties of these estimators are used to form recommendations for applied researchers.

Session 23INT51: Statistical Analysis of Streaming Data

Online Inference with Debiased Stochastic Gradient Descent

◆Ruijian Han[1], Lan Luo[2], Yuanyuan Lin[3] and Jian Huang[1]
[1]The Hong Kong Polytechnic University
[2]University of Iowa
[3]The Chinese University of Hong Kong

We propose a debiased stochastic gradient descent algorithm for online statistical inference with high-dimensional data. Our approach combines the debiasing technique developed in high-dimensional statistics with the stochastic gradient descent algorithm. It can be used for efficiently constructing confidence intervals in an online fashion. Our proposed algorithm has several appealing aspects: first, as a one-pass algorithm, it reduces the time complexity; in addition, each update step requires only the current data together with the previous estimate, which reduces the space complexity. We establish the asymptotic normality of the proposed estimator under mild conditions on the sparsity level of the parameter and the data distribution. We conduct numerical experiments to demonstrate the proposed debiased stochastic gradient descent algorithm reaches nominal coverage probability. Furthermore, we illustrate our method with a high-dimensional text dataset.

Inference in Heavy-Tailed Ar Models with Time Trends and Heteroscedastic Noises

Rui She
The Southwestern University of Finance and Economics

This paper studies estimation and inference in trend autoregression (AR) models with unspecified and heavy-tailed heteroscedastic noises. A piece-wise locally stationary structure of the noise is constructed to capture various heterogeneity and there is no restriction imposed on the tail index. The new nonstationary AR model allows for not only time- varying conditional features, but also unconditional variance and tail in- dex. This makes it appealing in practice, with wide applications in finance and econometrics. To obtain a feasible inference, we first investigate the self-weighted quantile regression estimator and derive its asymptotic nor- mality. Since the involved asymptotic variance depends on unobserved conditional density, a random weighting method is proposed to approx- imate the limiting distribution. The related Wald test is established to test the linear hypothesis in parameters. Based on residuals, a portman- teau test is further constructed to detect misspecifications in the fitted model. A simulation study and two applications to time series illustrate our inference procedure.

Irreversible Consumption Habit under Ambiguity: Singular Control and Optimal g-Stopping Time

Kyunghyun Park[1], ◆Kexin Chen[2] and Hoi Ying Wong[3]
[1]Nanyang Technological University
[2]The Hong Kong Polytechnic University
[3]The Chinese University of Hong Kong

Consider robust utility maximization with an irreversible consumption habit, where an agent concerned about model ambiguity is unwilling to decrease consumption and must simultaneously contend with a disutility (i.e., an adjustment cost) due to a consumption increase. While the optimization is a robust analog of singular control problems over a class of consumption-investment strategies and a set of probability measures, it is a new formulation that involves non-dominated probability measures of the diffusion process for the underlying assets in addition to singular controls with an adjustment cost. This paper provides a novel connection between the singular controls in the optimization and the optimal G-stopping times in a G-expectation space, using a duality theory. This connection enables us to derive the robust consumption strategy as a running maximum of the stochastic boundary, which is characterized by a free boundary arising from the optimal G-stopping times. The duality, which relies on arguments based on reflected G-BSDEs, is achieved by verifying the first-order optimality conditions for the singular control, the budget constraint equation for the robust strategies, and the worst-case realization under the non-dominated measures.

Blocked Gibbs Sampler for Truncated Two-Parameter Poisson-Dirichlet Process

◆Junyi Zhang[1] and Angelos Dassios[2]
[1]The Hong Kong Polytechnic University
[2]London School of Economics

The truncated two-parameter Poisson-Dirichlet process is a random probability measure, it provides a finite approximation to the distribution of the celebrated Pitman-Yor process. In this talk, we introduce the construction method and simulation algorithm of the truncated two-parameter Poisson-Dirichlet process. Then we show that such a process has a lower approximation error than the truncated stick-breaking process in approximating the Pitman-Yor process, and demonstrate its advantage in estimating the funtionals of the Pitman-Yor process. The truncated two-parameter Poisson-Dirichlet process implies a blocked Gibbs sampler that estimates the posterior of the Pitman-Yor process hierarchical models, we will illustrate this method with numerical examples.

Session 23INTSP1: Special Invited Session

Tba

Gang Li
UCLA
TBA

Semiparametric Predictive Inference for Failure Data using First-Hitting-Time Regression

◆Mei-Ling Ting Lee[1] and George Whitmore[2]
[1]University of Maryland
[2]McGill University

Degradation of an engineering system or disease progression in a patient can be described mathematically as a stochastic process. The system or the patient experiences a failure event when the wear and tear on the system or the patient's disease progression first reaches a critical threshold level. This happening defines a failure event and a first hitting time (FHT). First hitting time threshold regression (TR) models are based on an underlying stochastic process and hence the TR model represents a realistic alternative to the Cox model for capturing granular structure

in a prediction model. To date, most applications of threshold regression have been based on parametric families of stochastic processes. This paper presents a semiparametric form of threshold regression that requires the stochastic process to have only one key property, namely, stationary independent increments. Computational aspects of the methods are straightforward. We applied the methods to data from The Osteoarthritis Initiative (OAI) study are presented to demonstrate its practical use.

Session 23INT62: ML Meets Biostatistics: Theory and Practice

Supervised Knowledge may Hurt Novel Class Discovery Performance

♦Ziyun Li[1], Jona Otholt[1], Ben Dai[2], Di Hu[3], Christoph Meinel[1] and Haojin Yang[1]
[1]Hasso Plattner Institute
[2]Chinese University of Hong Kong
[3]Renmin University of China
Novel class discovery (NCD) aims to infer novel categories in an unlabeled dataset by leveraging prior knowledge of a labeled set comprising disjoint but related classes. Given that most existing literature focuses primarily on utilizing supervised knowledge from a labeled set at the methodology level, this paper considers the question: Is supervised knowledge always helpful at different levels of semantic relevance? To proceed, we first establish a novel metric, so-called transfer leakage, to measure the semantic similarity between labeled/unlabeled datasets. To show the validity of the proposed metric, we build up a large-scale benchmark with various degrees of semantic similarities between labeled/unlabeled datasets on ImageNet by leveraging its hierarachical class structure. The results based on the proposed benchmark show that the proposed transfer leakage is in line with the hierarachical class structure; and that NCD performance is consistent with the semantic similarities (measured by the proposed metric). Next, by using the proposed transfer leakage, we conduct various empirical experiments with different levels of semantic similarity, yielding that supervised knowledge may hurt NCD performance. Specifically, using supervised information from a low-similarity labeled set may lead to a suboptimal result as compared to using pure self-supervised knowledge. These results reveal the inadequacy of the existing NCD literature which usually assumes that supervised knowledge is beneficial. Finally, we develop a pseudo-version of the transfer leakage as a practical reference to decide if supervised knowledge should be used in NCD. Its effectiveness is supported by our empirical studies, which show that the pseudo transfer leakage (with or without supervised knowledge) is consistent with the corresponding accuracy based on various datasets.

De-Confounding Causal Inference using Latent Multiple-Mediator Pathways

♦Yubai Yuan[1] and Annie Qu[2]
[1]The Pennsylvania State University
[2]University of California, Irvine
Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the identification of both causal and mediation effects based on the proposed deconfounding method. Numerical experiments on both simulation settings and a normative aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

Causal Inference in Transcriptome-Wide Association Studies with Invalid Instruments and Gwas Summary Data

♦Haoran Xue, Xiaotong Shen and Wei Pan
University of Minnesota
Transcriptome-wide association studies (TWAS) have recently emerged as a popular tool to discover (putative) causal genes by integrating an outcome GWAS dataset with another gene expression/transcriptome GWAS (called eQTL) dataset. In our motivating and target application, we'd like to identify causal genes for low-density lipoprotein cholesterol (LDL), which is crucial for developing new treatments for hyperlipidemia and cardiovascular diseases. The statistical principle underlying TWAS is (two-sample) two-stage least squares (2SLS) using multiple correlated SNPs as instrumental variables (IVs); it is closely related to typical (two- sample) Mendelian randomization (MR) using independent SNPs as IVs, which is expected to be impractical and lower-powered for TWAS (and some other) applications. However, often some of the SNPs used may not be valid IVs, e.g. due to the widespread pleiotropy of their direct effects on the outcome not mediated through the gene of interest, leading to false conclusions by TWAS (or MR). Building on recent advances in sparse regression, we propose a robust and efficient inferential method to account for both hidden confounding and some invalid IVs via two-stage constrained maximum likelihood (2ScML), an extension of 2SLS. We first develop the proposed method with individual-level data, then extend it both theoretically and computationally to GWAS summary data for the most popular two-sample TWAS design, to which almost all existing robust IV regression methods are however not applicable. We show that the proposed method achieves asymptotically valid statistical inference on causal effects, demonstrating its wider applicability and superior finite-sample performance over the standard 2SLS/TWAS (and MR). We apply the methods to identify putative causal genes for LDL by integrating large-scale lipid GWAS summary data with eQTL data.

Knockofftrio: a Knockoff Framework for the Identification of Putative Causal Variants in Genome-Wide Association Studies with Trio Design

♦Yi Yang[1], Chen Wang[2], Linxi Liu[3], Joseph Buxbaum[4], Zihuai He[5] and Iuliana Ionita-Laza[2]
[1]City University of Hong Kong
[2]Columbia University
[3]University of Pittsburgh
[4]Icahn School of Medicine at Mount Sinai

[5]Stanford University

Family-based designs can eliminate confounding due to population substructure and can distinguish direct from indirect genetic effects, but these designs are underpowered due to limited sample sizes. Here, we propose KnockoffTrio, a statistical method to identify putative causal genetic variants for father-mother-child trio design built upon a recently developed knockoff framework in statistics. KnockoffTrio controls the false discovery rate (FDR) in the presence of arbitrary correlations among tests and is less conservative and thus more powerful than the conventional methods that control the family-wise error rate via Bonferroni correction. Furthermore, KnockoffTrio is not restricted to family-based association tests and can be used in conjunction with more powerful, potentially nonlinear models to improve the power of standard family-based tests. We show, using empirical simulations, that KnockoffTrio can prioritize causal variants over associations due to linkage disequilibrium and can provide protection against confounding due to population stratification. In applications to 14,200 trios from three study cohorts for autism spectrum disorders (ASDs), including AGP, SPARK, and SSC, we show that KnockoffTrio can identify multiple significant associations that are missed by conventional tests applied to the same data. In particular, we replicate known ASD association signals with variants in several genes such as MACROD2, NRXN1, PRKAR1B, CADM2, PCDH9, and DOCK4 and identify additional associations with variants in other genes including ARHGEF10, SLC28A1, ZNF589, and HINT1 at FDR 10%.

Session 23INT6: Bridging Statistics and Computation in High-Dimensional Data Analysis

Community Detection with Multiple Source of Information

♦Zongming Ma and Sagnik Nandy
University of Pennsylvania

In this talk, we discuss community detection when we observe m sparse networks and a high dimensional covariate matrix, all encoding the same community structure among n subjects. In the asymptotic regime where the number of features p and the number of subjects n grow proportionally, we derive an exact formula of asymptotic minimum mean square error (MMSE) for estimating the common community structure in the balanced two block case using an orchestrated approximate message passing algorithm. The formula implies the necessity of integrating information from multiple data sources. Consequently, it induces a sharp threshold of phase transition between the regime where detection (i.e., weak recovery) is possible and the regime where no procedure performs better than random guess. The asymptotic MMSE depends on the covariate signal-to-noise ratio in a more subtle way than the phase transition threshold. In the special case of $m = 1$, our asymptotic MMSE formula complements the pioneering work by Deshpande, Montanari, Mossel, and Sen, which found the sharp threshold when $m = 1$. A practical variant of the theoretically justified algorithm with spectral initialization leads to an estimator whose empirical MSEs closely approximate theoretical predictions over simulated examples.

Ranking Inferences Based on the Top Choice of Multiway Comparisons

Jianqing Fan[1], Zhipeng Lou[1], ♦Weichen Wang[2] and Mengxin Yu[1]

[1]Princeton University
[2]The University of Hong Kong

This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to M=2. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both ℓ_2-norm and ℓ_∞-norm, with the minimum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simulation studies lend further support to our theoretical results. A real data application illustrates the usefulness of the proposed methods convincingly.

using Svd for Topic Modeling

Tracy Ke
Harvard University

The probabilistic topic model imposes a low-rank structure on the expectation of the corpus matrix. Therefore, singular value decomposition (SVD) is a natural tool of dimension reduction. We propose an SVD-based method for estimating a topic model. Our method constructs an estimate of the topic matrix from only a few leading singular vectors of the corpus matrix, and has a great advantage in memory use and computational cost for large-scale corpora. The core ideas behind our method include a pre-SVD normalization to tackle severe word frequency heterogeneity, a post-SVD normalization to create a low-dimensional word embedding that manifests a simplex geometry, and a post-SVD procedure to construct an estimate of the topic matrix directly from the embedded word cloud. We provide the explicit rate of convergence of our method. We show that our method attains the optimal rate in the case of long and moderately long documents, and it improves the rates of existing methods in the case of short documents. The key of our analysis is a sharp row-wise large-deviation bound for empirical singular vectors, which is technically demanding to derive and potentially useful for other problems. We apply our method to a corpus of Associated Press news articles and a corpus of abstracts of statistical papers.

Approximate Message Passing from Random Initialization with Applications to Z2 Synchronization

♦Gen Li, Wei Fan and Yuting Wei
University of Pennsylvania

This talk is concerned with the problem of reconstructing an unknown rank-one matrix with prior structural information from noisy observations. While computing the Bayes-optimal estimator seems intractable in general due to its nonconvex nature, Approximate Message Passing (AMP) emerges as an efficient

first-order method to approximate the Bayes-optimal estimator. However, the theoretical underpinnings of AMP remain largely unavailable when it starts from random initialization, a scheme of critical practical utility. Focusing on a prototypical model called Z2 synchronization, we characterize the finite-sample dynamics of AMP from random initialization, uncovering its rapid global convergence. Our theory — which is non-asymptotic in nature — is the first in this model to unveil the non-necessity of a careful initialization for the success of AMP.

Session 23INT52: False Discovery Rate Control and Replicability Analysis of High Throughput Experiments

Statistical Assessment of Replicability: New Concepts and Methods

Xiaoquan Wen
University of Michigan

Statistical assessment of replicability is critical to ensure the quality and rigor of scientific research. In this talk, we discuss inference and modeling principles for replicability assessment. Targeting distinct application scenarios, we propose Bayesian model criticism and selection approaches to identify potentially irreplicable results in repeated scientific experiments. Our approaches are built upon established Bayesian prior and posterior predictive model-checking frameworks. They unify and generalize many existing replicability assessment methods. We discuss the application scenarios and statistical properties of the proposed methods and illustrate their usage by simulations and examples of real data analysis, including the data from the Reproducibility Project: Psychology and a systematic review of the impacts of pre-existing cardiovascular disease on COVID-19 outcomes.

An Innovative Nonparametric Procedure to Assess Reproducibility Across High-Throughput Studies

♦*Wen Zhou*[1], *Austin Ellingworth*[1], *Debashis Ghosh*[2] *and Zhigen Zhao*[3]
[1]Colorado State University
[2]University of Colorado Denver - Anschutz Medical Campus
[3]Temple University

Reproducibility is a fundamental aspect of experimental research that ensures the consistency and validity of findings. While there is a lack of consensus on how to assess reproducibility in general, in high-throughput studies, reproducibility is often defined as the consistency of test results across experiments. Most existing approaches either rely on stringent parametric assumptions of summary statistics or only focus on the hypothesis-wise alignment of summary statistics but overlook the experiment-wise heterogeneity. Inspired by Li et al. (2011) and Philtron et al. (2018), in this paper, we introduce a function based on the ranks of summary statistics from each experiment to define a notion for reproducibility and also to identify reproducible discoveries. The proposed nonparametric procedure takes into account both the signal strength and experiment-wise heterogeneity. Leveraging the lately developed concept about mirror statistics, we propose a novel procedure to identify reproducible findings using our procedure with marginal false discovery rate (mFDR) control. Our method controls the mFDR under fairly mild assumptions while outperforming existing methods in terms

of power. We validate the theoretical findings using comprehensive simulations and apply our method to two large-scale TWAS datasets to discover reproducible features. Overall, our proposed approach represents a significant advancement in the field of reproducibility research, with important practical implications for the identification of reproducible discoveries in high-throughput studies.

A New Fdr Controlling Procedure for Identifying Simultaneous Signals

Linsui Deng[1], *Kejun He*[1] *and* ♦*Xianyang Zhang*[2]
[1]Renmin University of China
[2]Texas A&M University

In many applications, identifying a single feature of interest requires testing the statistical significance of several hypotheses. Examples include mediation analysis which simultaneously examines the existence of the exposure-mediator and the mediator-outcome effects, and replicability analysis aiming to identify simultaneous signals that exhibit statistical significance across multiple independent experiments. In this work, we develop a novel FDR-controlling procedure, named the joint mirror (JM) procedure, to detect such features. The JM procedure exploits the mirror conservatism of the null p-values to construct a conservative estimate of the false discovery proportion (FDP). It iteratively shrinks the rejection region according to the partially revealed information until the FDP estimate is below the target level of false discovery rate (FDR). Adopting a Bayesian viewpoint, we develop an optimal rule for updating the rejection region and show that the resulting procedure controls the FDR in finite samples. We provide an efficient algorithm to implement the method. Extensive simulations demonstrate that our procedure can control the modified FDR, a more stringent error measure than the conventional FDR, and provide power improvement in several settings. Our method is further illustrated through real-world applications in mediation and replicability analyses.

Assessing Reproducibility of High-Throughput Experiments in the Case of Missing Data

Roopali Singh[1], *Feipeng Zhang*[2] *and* ♦*Qunhua Li*[1]
[1]Penn State University
[2]Xi'an JiaoTong Universitty

High-throughput experiments are an essential part of modern biological and biomedical research. The outcomes of high-throughput biological experiments often have a lot of missing observations due to signals below detection levels. For example, most single-cell RNA-seq (scRNA-seq) protocols experience high levels of dropout due to the small amount of starting material, leading to a majority of reported expression levels being zero. Though missing data contain information about reproducibility, they are often excluded in the reproducibility assessment, potentially generating misleading assessments. We develop a regression model to assess how the reproducibility of high-throughput experiments is affected by the choices of operational factors (eg, platform or sequencing depth) when a large number of measurements are missing. Using a latent variable approach, we extend correspondence curve regression, a recently proposed method for assessing the effects of operational factors to reproducibility, to incorporate missing values. We illustrate the usefulness of our method using a single-cell RNA-seq dataset collected on HCT116 cells. We compare the reproducibility of different library preparation platforms and study the effect of

sequencing depth on reproducibility, thereby determining the cost-effective sequencing depth that is required to achieve sufficient reproducibility.

Session 23INT81: Casual Inference in Biomedical Applications

The Synthetic Instrument: From Sparse Association to Sparse Causation

Dingke Tang, Dehan Kong and ♦Linbo Wang
University of Toronto

In many observational studies, researchers are interested in studying the effects of multiple exposures on the same outcome. Unmeasured confounding is a key challenge in these studies as it may bias the causal effect estimate. To mitigate the confounding bias, we introduce a novel device, called the synthetic instrument, to leverage the information contained in multiple exposures for causal effect identification and estimation. We show that under linear structural equation models, the problem of causal effect estimation can be formulated as an ℓ_0-penalization problem, and hence can be solved efficiently using off-the-shelf software. Simulations show that our approach outperforms state-of-art methods in both low-dimensional and high-dimensional settings. We further illustrate our method using a mouse obesity dataset.

Decision Trees with Fused Leaves for Prostate Cancer Diagnosis

Xiaogang Su
University of Texas at El Paso

We propose a new way of constructing decision trees, termed 'TreeFuL', which is short for 'trees with fused leaves'. TreeFuL combines recursive partitioning with fused regularization. Embedding a recursive partitioning step helps fused regularization with efficient grouping and interpretable results. At the same time, the fusion regularization naturally allows for amalgamation of non-neighboring terminal nodes, leading to more parsimonious final tree models. A cross-validation procedure is designed to incorporate both tree construction and tuning parameter selection. A leaf-coloring scheme is employed to complete tree shearing and node amalgamation in one step. We demonstrate the advantage and usage of the proposed method via both simulation studies and an application in prostate cancer diagnosis.

A Reference-Free r-Learner for Treatment Recommendation

♦Junyi Zhou[1], Ying Zhang[2] and Wanzhu Tu[3]
[1]Amgen Inc
[2]University of Nebraska Medical Center
[3]Indiana University-School of Medicine and Fairbanks School of Public Health

Assigning optimal treatments to individual patients based on their characteristics is the ultimate goal of precision medicine. Deriving evidence-based recommendations from observational data while considering the causal treatment effects and patient heterogeneity is a challenging task, especially in situations of multiple treatment options. Herein, we propose a reference-free R-learner based on a simplex algorithm for treatment recommendation. We showed through extensive simulation that the proposed method produced accurate recommendations that corresponded to optimal treatment outcomes, regardless of the

reference group. We used the method to analyze data from the Systolic Blood Pressure Intervention Trial (SPRINT) and achieved recommendations consistent with the current clinical guidelines.

Transporting Randomized Trial Results to Estimate Counterfactual Survival Functions in Target Populations

Zhiqiang Cao[1], ♦Youngjoo Cho[2] and Fan Li[3]
[1]Shenzhen Technology University
[2]Konkuk University
[3]Yale University

Generaliziability and transportability have been studied extensively for uncensored data. However, few literature of this topic focus on survival data with censoring. Motivated by controversial results from two clinical trials of blood pressure, in this article, we study transportability of survival outcome findings from randomized clinical trials to an external target population. Based on four assumptions, we propose inverse probability weighting estimators and doubly robust estimators and show that when both the sampling score model and censoring model are correctly specified, the proposed estimators are consistent. Furthermore, the doubly robust estimators are still consistent if the survival time model is correct no matter sampling score model and censoring model are misspecified or not. We derive influence functions of the proposed estimators and conduct simulation studies to examine their finite-sample performances. We finally apply our proposed estimators to assess transportability of survival difference between treatment and control groups found in ACCORD-BP trial to the adults with Diabetes mellitus of the U.S. population.

Session 23INT60: Stein's Method and Statistical Applications

Cramér-Type Moderate Deviation for Quadratic Forms with a Fast Rate

Xiao Fang[1], ♦Song-Hao Liu[2] and Qi-Man Shao[2]
[1]The Chinese University of Hong Kong
[2]Southern University of Science and Technology

Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors in $\mathbb{R}^d$. Suppose $\mathbb{E}X_1 = 0$, $\mathrm{Cov}(X_1) = I_d$, where I_d is the $d \times d$ identity matrix. Suppose further that there exist positive constants t_0 and c_0 such that $\mathbb{E}e^{t_0|X_1|} \leq c_0 < \infty$, where $|\cdot|$ denotes the Euclidean norm. Let $W = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$ and let Z be a d-dimensional standard normal random vector. Let Q be a $d \times d$ symmetric positive definite matrix whose largest eigenvalue is 1. We prove that for $0 \leq x \leq \varepsilon n^{1/6}$,

$$\left| \frac{\mathbb{P}(|Q^{1/2}W| > x)}{\mathbb{P}(|Q^{1/2}Z| > x)} - 1 \right| \leq C \left(\frac{1+x^5}{\det(Q^{1/2})n} 1_{\{d \geq 5\}} + \frac{1+x^3}{\det(Q^{1/2})n^{\frac{d}{d+1}}} 1_{\{d \leq 4\}} \right)$$

where ε and C are positive constants depending only on d, t_0, and c_0. This is a first extension of Cramér-type moderate deviation to the multivariate setting with a faster convergence rate than $1/\sqrt{n}$. The range of $x = o(n^{1/6})$ for the relative error to vanish and the dimension requirement $d \geq 5$ for the $1/n$ rate are both optimal. We prove our result using a new change of measure, a two-term Edgeworth expansion for the changed measure, and cancellation by symmetry for terms of the order $1/\sqrt{n}$.

Bounds for the Asymptotic Distribution of the Likelihood Ratio

♦*Andreas Anastasiou*[1] *and Gesine Reinert*[2]
[1]University of Cyprus
[2]University of Oxford

In this talk, we give an explicit bound on the distance to chi-square for the likelihood ratio statistic when the data are realisations of independent and identically distributed random elements. To our knowledge, this is the first explicit bound which is available in the literature. The bound depends on the number of samples as well as on the dimension of the parameter space. The bound is illustrated with three examples: samples from an exponential distribution, samples from a normal distribution and logistic regression.

Bootstrap Test for Multi-Scale Lead-Lag Relationships in High-Frequency Data

Takaki Hayashi[1] *and* ♦*Yuta Koike*[2]
[1]Keio University
[2]University of Tokyo

Motivated by recent empirical findings in high-frequency financial econometrics, we consider a pair of Brownian motions having possibly different lead-lag relationships at multiple time scales. Given their discrete observation data, we aim to test at which time scales these processes have non-zero cross correlations. For this purpose, we introduce maximum type test statistics based on scale-by-scale cross covariance estimators and develop a Gaussian approximation theory for these statistics. Since their null distributions are analytically intractable, we propose a wild bootstrap procedure to approximate them. Theoretical verification of these approximations are established through recent Gaussian approximation results for high-dimensional vectors of degenerate quadratic forms and based on the Malliavin-Stein method.

Cramér-Type Moderate Deviation for Sums of Local Dependent Random Variables

Song-Hao Liu and ♦*Zhuo-Song Zhang*
Southern University of Science and Technology
TBC

Session 23INT37: Bayesian Methods on Latent Variable Models

Bayesian Diagnostics of Hidden Markov Structural Equation Models with Missing Data

♦*Jingheng Cai*[1], *Ming Ouyang*[2], *Kai Kang*[1] *and Xinyuan Song*[3]
[1]Sun Yat-sen University
[2]The Chinese University of Hong Kong
[3]The Chinese University of Hong Kong,

Hidden Markov models (HMMs) are well suited to the characterization of longitudinal data in terms of a set of unobservable states, and have increasingly been used to uncover the dynamic heterogeneity in progressive diseases or activities. However, the existence of outliers or influential points may misidentify the hidden states and distort the associated inference. In this study, we develop a Bayesian local influence procedure for HMMs with latent variables in the presence of missing data. The proposed model enables us to investigate the dynamic heterogeneity of multivariate longitudinal data, reveal how the interrelationships among latent variables change from one state to another, and simultaneously conduct statistical diagnosis for the given data, model assumptions, and prior inputs. We apply the proposed procedure to analyze a dataset collected by the UCLA center for advancing longitudinal drug abuse research. Several outliers or influential points that seriously influence estimation results are identified and removed. The proposed procedure also discovers the effects of treatment and individuals'psychological problems on cocaine use behavior and delineates their dynamic changes across the cocaine-addiction states.

Variational Bayesian Inference for Two-Part Latent Variable Model

Yemao Xia

In this talk a variational Bayesian inference procedure is developed for the analysis of two-part latent variable model (TPLVM). By constructing a proper variational distribution family, the approximation to the posterior distribution is achieved via estimating variational parameters. We propose a scheme to update the variational parameters using the coordinate ascent inference (CAI) algorithm and develop a variational Bayes based procedure for the variable selection and model assessment. We conducted simulation studies to investigate the performance of our proposed method. Compared to the Markov Chains Monte Carlo (MCMC) sampling method, our proposed variational Bayesian (VB) approach achieves the computational efficiency without sacrificing estimation accuracy. We illustrate the practical merits of the VB approach by analyzing household finance survey data.

Latent Multiple Mediation Analysis with the Bayesian Lasso

Lijin Zhang[1] *and* ♦*Junhao Pan*[2]
[1]Graduate School of Education, Stanford University, U.S.
[2]Department of Psychology, Sun Yat-sen University, Guangzhou, China

Mediators have played an essential role in helping researchers understand the mechanism through which the predictors affect the outcome variables. The existence of multiple mediators is also very common in behavior research. Traditional approaches for testing the indirect effects include the Sobel test, percentile bootstrap method, and bias-corrected bootstrap method. When handling multiple mediators simultaneously, the traditional approaches for testing the indirect effects, which include the Sobel test and bootstrap method, are prone to inflated Type I error rates and the overfitting problem. To provide a more effective variable selection tool in multiple mediation analysis, Serang et al. (2017) integrated mediation models of observed variables with the frequentist Lasso (least absolute shrinkage and selection operator) method. However, this method has two limitations: (1) it doesn't take the measurement errors of manifest variables into account; (2) it cannot provide uncertainty information about indirect effects (e.g., interval estimation). The current study extended the Bayesian Lasso method into the framework of latent multiple mediation models to solve the above-mentioned problems. A Monte Carlo simulation study was conducted to compare the proposed method with the traditional Sobel and bootstrap methods. Recommendations and future directions were also provided based on the findings of the simulation study.

Joint Analysis of Mixed Types of Outcomes with Latent Variables

♦*Deng Pan*[1], *Yingying Wei*[2] *and Xinyuan Song*[2]

[1]School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China
[2]Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong

We propose a joint modeling approach to investigating the observed and latent risk factors of mixed types of outcomes. The proposed model comprises three parts. The first part is an exploratory factor analysis model that summarizes latent factors through multiple observed variables. The second part is a proportional hazards model that examines the observed and latent risk factors of multivariate time-to-event outcomes. The third part is a linear regression model that investigates the determinants of a continuous outcome. We develop a Bayesian approach coupled with MCMC methods to determine the number of latent factors, the association between latent and observed variables, and the important risk factors of different types of outcomes. A modified stochastic search item selection algorithm, which introduces normal-mixture-inverse gamma priors to factor loadings and regression coefficients, is developed for simultaneous model selection and parameter estimation. The proposed method is subjected to simulation studies for empirical performance assessment and then applied to a study concerning the risk factors of type 2 diabetes and the associated complications.

Session 23INT80: Recent Developments in Survival Analysis

Health Utility Survival for Randomized Clinical Trials: Extensions and Statistical Properties

Yangqing Deng[1], ♦Meiling Hao[2], John R. De Lmeida[3] and Wei Xu[4]
[1]Princess Margaret Cancer Centre, University Health Network,
[2]University of International Business and Economics
[3]Department of Otolaryngology—H&N Surgery, University Health Network
[4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Overall survival has been used as the primary endpoint for many randomized trials that aim to examine whether a new treatment is non-inferior to the standard treatment or placebo control. When a new treatment is indeed non-inferior in terms of survival, it may be important to assess other outcomes including health utility. However, analyzing health utility scores in a secondary analysis may have limited power since the primary objectives of the original study design may not include health utility. To comprehensively consider both survival and health utility, we developed a composite endpoint, HUS (Health Utility-adjusted Survival), which combines both survival and utility. With a detailed framework to conduct sample size calculation and power analysis, HUS has been shown to be able to increase statistical power and potentially reduce the required sample size compared to the standard overall survival endpoint. Nevertheless, the asymptotic properties of the test statistics of HUS endpoint have yet to be fully established. Besides that, the standard version of HUS cannot be applied to or have limited performance in certain scenarios, where extensions are needed. In this manuscript, we propose various methodological extensions of HUS and derive the asymptotic distributions of the test statistics. By comprehensive simulation studies and a data applica-

tion using retrospective data based on a translational patient cohort in Princess Margaret Cancer Centre, we demonstrate the better efficiency and feasibility of HUS comparing to different methods.

Transfer Learning by Optimal Model Averaging for Censored Data

♦*Baihua He[1] and Xinyu Zhang[2]*
[1]University of Science and Technology of China
[2]Chinese Academy of Sciences

Transferring information plays an important role in summarizing and synthesizing scientific results derived from multiple datasets to predict clinical outcomes. When the main research model of interest constructed by one local dataset shares covariate effects with other data sources, incorporating information from these datasets can improve prediction accuracy. Existing methods require correct model specifications, which may not be practical, especially with multiple dataset sources. We develop a transfer learning approach by model averaging for predicting censored responses. Specifically, several helper models are formulated with shared parameters from other datasets, and the optimal weights for the averaging procedure are derived by minimizing a delete-one cross-validation criterion. The proposed method allows the model framework to vary among helper models. We show that the proposed approach asymptotically achieves the lowest prediction risk if the main model is misspecified. In addition, the proposed procedure attains model weight consistency if the main model is correctly specified. We further demonstrate that the risk of the proposed model averaging approach is no larger than the risks of the equal weighting approach and the pure model selection asymptotically, regardless of whether the main model is correct. We conduct extensive numerical studies to demonstrate the superior performances of the proposed procedure over the other existing methods, and we further show this using four gene expression dataset sources for patients with breast cancer.

Simultaneous Variable Selection and Estimation for Interval-Censored Failure Time Data with Ancillary Information

♦*Mingyue Du[1] and Xingqiu Zhao[2]*
[1]The Hong Kong Polytechnic University Shenzhen Research Institute
[2]The Hong Kong Polytechnic University

Simultaneous variable selection and estimation has recently attracted a great deal of attention and in particular, many methods have been proposed for it in the context of failure time data. We consider the same problem but for a situation that has been not discussed and when one faces interval-censored data with the presence of some ancillary information given in the form of recurrent event processes. One special case of such situations is interval-censored data with informative censoring. For the problem, a conditional proportional hazards model is proposed and a penalized sieve maximum likelihood procedure is developed for the simultaneous variable selection and estimation of covariate effects. In the method, B-splines functions are used and the oracle property of the proposed estimators is established. A simulation study is conducted to assess the finite sample performance of the proposed approach and suggests that it works well. The method is applied to a set of real data arising from an Alzheimer's disease study.

Session 23INT78: Recent Advances in Long-Run Variance Estimation in Time Series and Spatial Data

Variance Estimation of Spatial Autocorrelated Data under Non-Constant Mean

♦Di Su and Kin Wai Chan
The Chinese University of Hong Kong

Asymptotic variance estimation plays an essential role in various inference problems, and in practice, the underlying process is usually masked by some mean structure making the problem difficult. For one-dimensionally-indexed data, the difference sequence method has been studied to avoid estimating the mean function. However, when the sampling space becomes two-dimensionally-indexed, difference sequence methods have only been studied for independent data. To fill in this gap, we provide a class of difference-based estimators of asymptotic variance for spatial autocorrelated data. We show that when the spatial data are autocorrelated, the optimal spatial difference sequences are different from those used in time series literature, and we provide their numerical values. We find that unless the mean function satisfies certain small regions assumption, the difference sequence method may not perform as well as if there was no mean effect. However, it still performs better than its non-difference-based counterpart. The optimal bandwidths are provided and the estimator's performance is demonstrated through numerical studies. An R package daves is available for its implementation.

Revamping Kernel-Based Long-Run Variance Estimation: a Converging Kernel Approach

♦Xu Liu and Kin Wai Chan
The Chinese University of Hong Kong

Kernel estimators have been popular for decades in long-run variance estimation. To minimize the loss of efficiency measured by the mean-squared error in important aspects of kernel estimation, we propose a novel class of converging kernel estimators that have the ño-loseproperties including: (1) no efficiency loss from estimating the bandwidth as the optimal choice is universal; (2) no efficiency loss from ensuring positive-definiteness using a principle-driven aggregation technique; and (3) no efficiency loss asymptotically from potentially misspecified prewhitening models and transformations of the time series. A shrinkage prewhitening transformation is proposed for more robust finite-sample performance. The estimator has a positive bias that diminishes with the sample size so that it is more conservative compared with the typically negatively biased classical estimators. The proposal improves upon all standard kernel functions and can be well generalized to the multivariate case. We discuss its performance through simulation results and one real-data application % including the forecast breakdown test and in MCMC convergence diagnostics.

Reorder-Based Variance Estimator in Nonparametric Regression with Time Series Errors

♦Zerun Wang and Kin Wai Chan
The Chinese University of Hong Kong
TBC

Recursive Nonparametric Estimation: Principles, Methods and Applications

♦Man Fung Leung[1] and Kin Wai Chan[2]
[1]University of Illinois Urbana-Champaign
[2]The Chinese University of Hong Kong

Existing long-run variance estimators face a dilemma between mean squared error, time complexity, and space complexity. In this talk, we will present a conceptual decomposition to understand this phenomenon. The new insights allow us to improve existing works, but we further characterize efficient estimators in a principle-driven way. Our asymptotic theory and simulations show that this methodological approach leads to online estimators with a lower mean squared error. We also discuss practical enhancements such as mini-batch and automatic updates. Encouraging finite-sample results are illustrated in online change point detection, stochastic approximation, and Markov chain Monte Carlo convergence diagnosis.

Session 23INT107: Recent Advances in Statistical Methods for Analyzing High-Dimensional Cancer and Disease Surveillance Data

High Dimensional Gaussian Graphical Regression Models with Covariates

Jingfei Zhang[1] and ♦Yi Li[2]
[1]Univ of Miami
[2]Univ of Michigan

Though Gaussian graphical models have been widely used in many scientific fields, relatively limited progress has been made to link graph structures to external covariates. We propose a Gaussian graphical regression model, which regresses both the mean and the precision matrix of a Gaussian graphical model on covariates. In the context of co-expression quantitative trait locus (QTL) studies, our method can determine how genetic variants and clinical conditions modulate the subject-level network structures, and recover both the population-level and subject-level gene networks. Our framework encourages sparsity of covariate effects on both the mean and the precision matrix. In particular for the precision matrix, we stipulate simultaneous sparsity, i.e., group sparsity and element-wise sparsity, on effective covariates and their effects on network edges, respectively. We establish variable selection consistency first under the case with known mean parameters and then a more challenging case with unknown means depending on external covariates, and establish in both cases the convergence rates of the estimated precision parameters. The utility and efficacy of our proposed method is demonstrated through simulation studies and an application to a co-expression QTL study with brain cancer patients.

Approximate Bayesian Computation Estimation of Models for the Natural History of Breast Cancer, with Application to Data from a Milan Cohort Study

♦Marco Bonetti[1], Laura Bondi[2], Denitsa Grigorova[3] and Antonio Russo[4]
[1]Bocconi University, Milan, Italy
[2]Cambridge University, Cambridge, UK
[3]Sofia University, Sofia, Bulgaria
[4]UOC Osservatorio Epidemiologico, ATS, Milan, Italy

We develop multi-state models for the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease and the start of symptomatic detectability. The former kind of detection occurs through screening, while the latter through the insurgence of symptoms.

We develop a cure rate parametric specification that allows for dependence between the times from birth to the two events, and present the results of the analysis of data collected as part of a motivating study from Milan. Participants in the study had a varying degree of compliance to a regional breast cancer screening program. The subjects' ten-year trajectories have been obtained from administrative data collection performed by the Italian national health care system. We first present a tractable model for which we develop the likelihood contributions of the possible observed trajectories, and perform maximum likelihood inference on the latent process. We rely on a likelihood-free method, Approximate Bayesian Computation (ABC), for inference on such more flexible models. Issues that arise from the use of ABC for model choice and parameter estimation are discussed, with a focus on the problem of choosing appropriate summary statistics. The estimated parameters of the underlying disease process allow for the study of the effect of different examination schedules (ages and frequencies for screening examinations) and different adherence patterns on a population of asymptomatic subjects. We also discuss extensions to these models and specific issues associated with the incorporation of family history information in such models.

New Statistical Method for Spatio-Temporal Surveillance of Infectious Diseases

Peihua Qiu
University of Florida

Online sequential monitoring of the incidence rates of infectious diseases is critically important for public health. Governments have invested a great amount of resource in building global, national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) charts, are routinely included for disease surveillance purposes. However, these charts require many assumptions on the observed data that are rarely valid in practice, making their results unreliable. In this talk, we present a new sequential monitoring approach for spatio-temporal disease surveillance, which can accommodate the dynamic nature of the observed disease incidence rates, spatio-temporal data correlation, and nonparametric data distribution. It is shown that the new method is more reliable to use in practice than the commonly used conventional control charts for spatio-temporal surveillance of infectious diseases.

Session 23INT42: Recent Advances and Applications of Survival Analysis in Biomedical Research

Cox Proportional Hazards Regression with Interval Censored Outcome and Covariate

♦*Dongdong Li[1], Yue Song[2], Wenbin Lu[3], Huldrych Gunthard[4], Roger Kouyos[4] and Rui Wang[1]*
[1]Harvard Medical School
[2]Harvard School of Public Health
[3]North Carolina State University
[4]University of Zurich

In HIV cure research, understanding predictors of viral rebound trajectories after antiretroviral therapy (ART) interruption is crucial. We examine modeling approaches connecting interval-censored outcomes and covariates, like time to viral rebound and suppression. We develop a proportional hazards regression model for interval-censored outcome and covariate estimation and inference without making parametric assumptions about baseline hazard functions, using an Expectation-Maximization algorithm. The method is extended to accommodate multiple ART initiation and interruption episodes from the same participant. We assess the proposed method's finite-sample performance in independent and clustered data settings via simulation studies and evaluate the effect of time to viral suppression.

Recent Advances in Handling Time-to-Event Data with Internal Covariates

♦*Trevor Thomson, Joan Hu and Bohdan Nosyk*
Simon Fraser University

Previous studies indicate that retention on an opioid agonist treatment (OAT) can reduce the mortality risk of people with opioid use disorder. To account for the dynamic nature of OAT use, we considered an extended Cox proportional hazards model to account for treatment history through a time-dependent stratification variable. As OAT use is a time-dependent internal covariate, conventional likelihood / partial likelihood based methods do not directly apply, and an estimating function based procedure for estimating model parameters is proposed. As the model makes explicit use of the time-dependent internal covariate, estimating survival probabilities based on the model is no longer feasible. With the aim of obtaining such probabilities, we considered an alternative Cox proportional hazards model, where the internal covariate is replaced with a latent random variable, in which its distribution depends on the entire history of the covariate process. Upon modelling the OAT usage process, we predict the latent variable based on an individual's observed history, and use the conditional score approach to derive unbiased estimating equations for the parameters of interest. The resulting procedure serves as an alternative to traditional joint modelling that requires strong parametric assumptions, and can be computationally intensive. The methods are illustrated using an administrative dataset capturing OAT dispensations and deaths in British Columbia, Canada. This is joint work with Joan Hu (SFU) and Bohdan Nosyk (St. Paul's Hospital and SFU).

Deep Neural Network with a Smooth Monotonic Output Layer for Dynamic Risk Prediction

Zhiyang Zhou
University of Manitoba

Risk prediction is a key component of survival analysis in medicine, public health, economics, engineering, and many other areas. The fundamental concern of risk prediction is the relationship between predictors and the survival function. The recent success of survival analysis has already been extended to dynamic risk prediction, where the model considers repeated measurements of time-varying predictors. However, existing approaches usually involve strong model assumptions or discretize the time domain, which may lead to biased prediction. To tackle these issues, we present a deep neural network with a novel output layer termed the Smooth Monotonic Output Layer. The resulting network involves no discretization and specifies no parametric structure for the underlying relationship between predictors and the time to event. Extensive results show that our proposal achieves state-of-the-art accuracy in predicting the individual-level risk of atherosclerotic cardiovascular disease.

Session 23INT64: Statistical Methods in Data Integration and Synthesis

Repro Samples Method for Uncertainty Quantification in Irregular Inference Problems and more

Minge Xie
Rutgers University

Increasingly more complex and diverse problems in data science demand us to have new inferential frameworks to tackle highly non-trivial irregular inference problems, for instance, those involving discrete or non-numerical parameters and those involving non-numerical data, etc. In this talk, we present a novel, wide-reaching and effective simulation-inspired framework, called repro samples method, to conduct statistical inference and quantify uncertainty for the irregular problems plus more. We systemically develop both exact and approximate (asymptotic) theories to support the development. An attractive feature is that the development doesn't need to rely on a likelihood or use the large sample central limit theorem, and thus is especially effective for complicated and irregular inference problems often encountered in machine learning and data science. The effectiveness of the proposed method is illustrated through two case study examples that provide solutions to two open inference problems in statistics: (a) In a Gaussian mixture model, how to construct a confidence set for the unknown number of components? (b) In a high dimensional regression model, how to construct confidence sets for either the unknown true model, a single or a collection of regression coefficients, or both model and regression coefficients jointly? Comparison studies demonstrate that the proposed methods have far superior performance to existing Bayesian and frequentist attempts. Although the case studies pertain to the traditional statistics models, the framework also has direct extensions to other irregular inference problems and complex machine learning models, such as rare-events meta-analysis, graphical network, random forest, voice data, etc.

Meta-Analysis of Safety Data

Shouhao Zhou
Penn State University

Meta-analysis is a powerful tool for drug safety assessment by synthesizing treatment-related toxicological findings from independent clinical trials. However, published clinical studies may or may not report all adverse events (AEs) if the observed number of AEs were fewer than a pre-specified study-dependent cutoff. Subsequently, with censored information ignored, the estimated incidence rate of AEs could be significantly biased. To address this common problem in meta-analysis, we propose a general Bayesian multilevel regression model to accommodate the censored rare event data, and investigate its performance to identify high-risk subgroups. The proposed approach is illustrated using data from a meta-analysis of 125 clinical trials involving PD-1/PD-L1 inhibitors with respect to their toxicity profiles.

A Unifying Dependent Combination Framework with Applications to Association Tests

♦*Xiufan Yu*[1], *Linjun Zhang*[2], *Arun Srinivasan*[3], *Lingzhou Xue*[4]
and Minge Xie[2]
[1]University of Notre Dame
[2]Rutgers University
[3]GSK plc
[4]Penn State University

In this work, we propose a generic framework to combine arbitrarily dependent p-values with statistical guarantees. Under the framework, we make a connection to the conventional meta-analysis methods of combining p-values from independent sources and provide a comprehensive study on various dependent combination methods. In particular, we show that the popular Cauchy combination method can be regarded as a special case within this framework. Moreover, the proposed framework provides a way to modify the Cauchy combination when the distributional assumptions for the standard Cauchy method are violated. As two illustrative examples, we apply the proposed framework to construct dependence-adjusted combined association tests in genetic studies and microbiome research. Through the aggregation of information from multiple existing tests, the combined tests leverage the respective strengths of each existing test, and the testing power is substantially boosted toward a wide range of alternative spaces. Our numerical results demonstrate that ignoring the dependence among the to-be-combined components may lead to a severe size distortion phenomenon. Compared to traditional independence-assumed methods, the proposed combination framework can handle the dependence accurately and utilizes the information efficiently to construct tests with accurate size and enhanced power.

Optimizing Information Borrowing for Bayesian Hierarchical Model in Subgroup Analysis

Xuetao Lu and ♦*J. Jack Lee*

University of Texas MD Anderson Cancer Center

Subgroups occur naturally in a large variety of data sets and data analysis. For example, how do we estimate the efficacy of immunotherapy in different molecular subtypes of lung cancer? How do we estimate the 30-day mortality rate of COVID infection in different countries? Bayesian hierarchical model (BHM) has been widely used in synthesizing information across subgroups. The typical assumption of exchangeability is very restricted and often does not hold. Efforts have been made in clustering the subgroups first, then, assuming exchangeability within cluster and borrowing information across subgroups within the same cluster. The two-step procedure has two main challenges: (1) How to determine the number of clusters? And (2) How much information to borrow within each cluster? To address these two interconnected challenges, we propose two distribution-free overlapping indices, namely, the overlapping clustering index for identifying the optimal clustering result and the overlapping borrowing index for assigning proper borrowing strength to clusters. Accordingly, we develop a new method BHMOI (Bayesian hierarchical model with overlapping indices). BHMOI includes a novel weighted K-Means clustering algorithm to obtain optimal clustering results, and an innate way to dynamically determining the borrowing strength in each cluster. BHMOI can achieve efficient and robust information borrowing with desirable properties. Examples and simulation studies are provided to demonstrate the effectiveness of BHMOI in heterogeneity identification and dynamic information borrowing.

Session 23INT4: Recent Development on Analysis of Complex Time-to-Event Data

Feature Screening with Large Scale and High Dimensional Survival Data

Grace Yi[1], *♦Wenqing He*[1] *and Raymond Carroll*[2]
[1]University of Western Ontario
[2]Texas A&M University and University of Technology Sydney

Data with a huge size present great challenges in modeling, inferences, and computation. In handling big data, much attention has been directed to settings with "large   small ", and relatively less work has been done to address problems with   and   being both large, though data with such a feature have now become more accessible than before, where   represents the number of variables and   stands for the sample size. The big volume of data does not automatically ensure good quality of inferences because a large number of unimportant variables may be collected in the process of gathering informative variables. To carry out valid statistical analysis, it is imperative to screen out noisy variables that have no predictive value for explaining the outcome variable. In this paper, we develop a screening method for handling large-sized survival data, where the sample size is large and the dimension   of covariates is of non-polynomial order of the sample size . We rigorously establish theoretical results for the proposed method and conduct numerical studies to assess its performance. Our research offers multiple extensions of existing work and enlarges the scope of high-dimensional data analysis. The proposed method capitalizes on the connections among useful regression settings and offers a computationally efficient screening procedure. Our method can be applied to different situations with large-scale data including genomic data.

Linearized Maximum Rank Correlation Estimation

Guohao Shen[1], *Kani Chen*[2], *Jian Huang*[1] *and ♦Yuanyuan Lin*[3]
[1]The Hong Kong Polytechnic University
[2]Hong Kong University of Science and Technology
[3]The Chinese University of Hong Kong

We propose a linearized maximum rank correlation estimator for the single-index model. Unlike the existing maximum rank correlation and other rank-based methods, the proposed estimator has a closed-form expression, making it appealing in theory and computation. The proposed estimator is robust to outliers in the response and its construction does not need knowledge of the unknown link function or the error distribution. Under mild conditions, it is shown to be consistent and asymptotically normal when the predictors satisfy the linearity of the expectation assumption. A more general class of estimators is also studied. Inference procedures based on the plug-in rule or random weighting resampling are employed for variance estimation. The proposed method can be easily modified to accommodate censored data. It can also be extended to deal with high-dimensional data combined with a penalty function. Extensive simulation studies provide strong evidence that the proposed method works well in various practical situations. Its application is illustrated with the Beijing PM 2.5 dataset.

Efficient Estimation for the Accelerated Failure Time Model with Auxiliary Aggregated Information

♦Huijuan Ma[1], *Yukun Liu*[1], *Donglin Zeng*[2] *and Yong Zhou*[1]
[1]East China Normal University
[2]University of North Carolina

With the rapidly increasing availability of aggregated data in the public domain, there has been a growing interest in synthesizing information from individual-level data and aggregated data. This article proposes a novel one-step estimator to improve the estimation of the accelerated failure time model by incorporating the subgroup survival probabilities from external sources. The proposed framework stems from the maximum full likelihood estimator and allows for population heterogeneity. We establish the consistency and asymptotic normality of the proposed estimator and show that it is more efficient than the maximum conditional likelihood estimator without combining information. The asymptotic variance of the proposed estimator has a closed form and its variance estimator is easily obtained by plug-in rule. Simulation studies show that the proposed estimator yields an efficiency gain over existing approaches. The proposed methodology is illustrated with an analysis of a chemotherapy study for Stage III colon cancer.

Session 23INT24: Recent Advances in Nonparametric Statistics and Novel Applications

Variable Selection in Semiparametric Transformation Regression with Interval-Censored Competing Risks Data

Fatemeh Mahmoudi and ♦Xuewen Lu
University of Calgary

In the framework of variable selection problems, penalized regression is a popular approach. Although there have been numerous types of research on penalized variable selection methods for standard time-to-event data and regression models, they are not applicable when data are interval-censored competing risks. In this context, we develop a penalized variable selection procedure that is able to handle such data in a broad class of semiparametric transformation regression models, which contain some popular models such as the proportional and non-proportional hazards models as special cases and allow for direct assessment of covariate effects on the cumulative incidence or sub-distribution function of competing risks. The proposed penalized variable selection strategy can simultaneously handle variable selection and parameter estimation. We rigorously establish the asymptotic properties of the proposed penalized estimators and modify the EM algorithm and coordinate descent algorithm for implementation. Simulation studies are conducted to demonstrate the good performance of the proposed method. Some real data examples are used for illustration.

Optimal-k Sequence for Difference-Based Methods in Nonparametric Regression

Wenlin Dai[1], *Xingwei Tong*[2] *and ♦Tiejun Tong*[3]
[1]Renmin University of China
[2]Beijing Normal University
[3]Hong Kong Baptist University

Difference-based methods have been attracting increasing attention in nonparametric regression, in particular for estimating the residual variance. To implement the estimation, one needs to choose an appropriate difference sequence, mainly between the optimal difference sequence and the ordinary difference sequence. The difference sequence selection is a fundamental problem in nonparametric regression, and it remains a controversial issue for over three decades. In this paper, we propose to tackle this challenging issue from a very unique perspective, namely by introducing a new difference sequence called the

optimal-k difference sequence. The new difference sequence not only provides a better balance between the bias-variance trade-off, but also dramatically enlarges the existing family of difference sequences that includes the optimal and ordinary difference sequences as two important special cases. We further demonstrate, by both theoretical and numerical studies, that the optimal-k difference sequence has been pushing the boundaries of our knowledge in difference-based methods in nonparametric regression, and it always performs the best in practical situations.

Hypotheses Testing of Functional Principal Components

Zening Song[1], ♦Lijian Yang[2] and Yuanyuan Zhang[3]

[1]Nankai University

[2]Tsinghua University

[3]Soochow University

We propose a test for the hypothesis that the standardized functional principal components (FPCs) of a functional data equal a given set of orthonormal basis (e.g., the Fourier basis). Using estimates of individual trajectories that satisfy certain approximation conditions, a chi-square type statistic is constructed and shown to be oracally efficient under the null hypothesis in the sense that its limiting distribution is the same as an infeasible statistic using all trajectories, known by "oracle". The null limiting distribution is an infinite Gaussian quadratic form, and a consistent estimator of its quantile is obtained. A test statistic based on the chi-square type statistic and approximate quantile of the Gaussian quadratic form is shown to be both of the nominal asymptotic significance level and asymptotically correct. It is further shown that B-spline trajectory estimates meet the required approximation conditions. Simulation studies illustrate superior finite sample performance of the proposed testing procedure. For the EEG (ElectroEncephalogram) data, the proposed procedure has confirmed an interesting discovery that the centered EEG data is generated from a small number of elements of the standard Fourier basis.

A Nearest Neighbor Method for Continuous Stochastic Optimization

♦Seksan Kiatsupaibul[1], Pariyakorn Maneekul[2] and Zelda Zabinsky[2]

[1]Chulalongkorn University

[2]University of Washington

We propose a version of the nearest neighbor method in machine learning that can be used as a global optimization algorithm for continuous stochastic optimization. The proposed algorithm is an adaptation of the single observation search algorithm (SOSA) that incorporates the upper confidence bound approach into its sampling process and adopts the quadratic function approximation for its function value estimation. The upper confidence bound component balances exploration and exploitation of SOSA, enhancing the robustness of the sampling process against random errors. The quadratic function approximation offers a more accurate generalization of the function values within the neighborhood of an optimal solution, increasing the accuracy of the optimal value estimation. The numerical experiments illustrate the behavior of the proposed method when it is applied to well-known test functions. The results show that the proposed method enhances the performance of SOSA in terms of both accuracy and robustness.

Session 23INT39: Recent Developments on Complex Data Analysis

Earthquake Parametric Insurance with Bayesian Spatial Quantile Regression

Jefferey Pai[1], ♦Yunxian Li[2], Aijun Yang[3] and Chenxu Li[4]

[1]University of Manitoba

[2]University of Yunnan University of Finance and Economics

[3]Nanjing Forest University

[4]Yunnan University of Finance and Economics

Earthquake Parametric Insurance with Bayesian Spatial Quantile Regression   With its transparent and fast claims payment, parametric insurance has been widely used to insure nature-related risks such as earthquakes, floods, and hurricanes. In 2014, earthquake parametric insurance was introduced to provide coverage for earthquake losses occurred in Yunnan Province of China. However, as a main limitation of parametric insurance, basis risk is inevitable. In this paper, a Bayesian spatial quantile regression model is proposed to reduce the basis risk of earthquake parametric insurance. The effect of earthquake hazard, risk exposure, and vulnerability on economic loss are analyzed and considered in the quantile regression model. Since risk exposure and vulnerability at the epicenter cannot be observed, they will be treated as latent variables in the quantile regression model. Bayesian approaches are applied, and spatial correlation is considered to construct the prior distributions for the latent variables. Earthquake losses in Yunnan Province from 1992 to 2019 are collected and analyzed by the proposed model and methods. The payment mechanism and the corresponding premiums of 16 regions in Yunnan Province are then calculated. The results show that the loss ratio is more reasonable than the current earthquake insurance, and the basis risk is then reduced.

Optimal Integrating Learning for Split Questionnaire Design Type Data

Cunjie Lin[1], Jingfu Peng[1], Yichen Qin[2], ♦Yang Li[1] and Yuhong Yang[3]

[1]Renmin University of China

[2]University of Cincinnati

[3]University of Minnesota

In the era of data science, it is common to encounter data with different subsets of variables obtained for different cases. An example is the split questionnaire design (SQD), which is adopted to reduce respondent fatigue and improve response rates by assigning different subsets of the questionnaire to different sampled respondents. A general question then is how to estimate the regression function based on such block-wise observed data. Currently, this is often carried out with the aid of missing data methods, which may unfortunately suffer from intensive computational cost, high variability, and possible large modeling biases in real applications. In this article, we develop a novel approach for estimating the regression function for SQD-type data. We first construct a list of candidate models using available data-blocks separately, and then combine the estimates properly to make an efficient use of all the information. We show the resulting averaged model is asymptotically optimal in the sense that the squared loss and risk are asymptotically equivalent to those of the best but infeasible averaged estimator. Both simulated examples and an application to the SQD dataset from the European Social Survey show the promise of the proposed method.

Robust Estimation and Test Based on Median-of-Means Method

Pengfei Liu
Jiangsu Normal University

Using the idea of grouping under moderate data framework, we propose the median-of-means(MoM) type nonparametric estimator for parameters of statistical model which has been used widely in various disciplines. Under certain condition on the growing rate of the number of subgroups, the consistency and asymptotic normality of the proposed estimator are investigated. Furthermore, we construct a new method to test the parameters based on the empirical likelihood method for median. Extensively numerical simulations are designed to demonstrate the superiorities of our estimator. It is shown that the new proposed estimator is quite robust with respect to outliers. We also apply the MoM method to analyze some real data sets.

Additive Hazards Model with Time-Varying Coefficients and Imaging Predictors

♦*Qi Yang[1], Chuchu Wang[2], Haijin He[3], Xiaoxiao Zhou[2] and Xinyuan Song[2]*
[1]School of Management, Shandong University
[2]The Chinese University of Hong Kong
[3]Shenzhen University

Conventional hazard regression analyses frequently assume constant regression coefficients and scalar covariates. However, some covariate effects may vary with time. Moreover, medical imaging has become an increasingly important tool in screening, diagnosis, and prognosis of various diseases, given its information visualization and quantitative assessment. This study considers an additive hazards model with time-varying coefficients and imaging predictors to examine the dynamic effects of potential scalar and imaging risk factors for the failure of interest. We develop a two-stage approach that comprises the high-dimensional functional principal component analysis technique in the first stage and the counting process-based estimating equation approach in the second stage. In addition, we construct the pointwise confidence intervals for the proposed estimators and provide a significance test for the effects of scalar and imaging covariates. Simulation studies demonstrate the satisfactory performance of the proposed method. An application to the Alzheimer's disease neuroimaging initiative study further illustrates the utility of the methodology.

Session 23INT89: Challenges and Developments in Econometrics and Statistical Theories

Spiked Eigenvalues of High-Dimensional Sample Autocovariance Matrices: Clt and Applications

Daning Bi[1], Xiao Han[2], Adam Nie[1] and ♦Yanrong Yang[1]
[1]Australian National University
[2]University of Science and Technology of China

High-dimensional autocovariance matrices play an important role in dimension reduction for high-dimensional time series. In this article, we establish the central limit theorem (CLT) for spiked eigenvalues of high-dimensional sample autocovari- ance matrices, which are developed under general conditions. The spiked eigenvalues are allowed to go to infinity in a flexible way without restrictions in divergence order. Moreover, the number of spiked eigenvalues and the time lag of the autocovariance matrix under this study could be either fixed or tending to infinity when the di- mension p and the time length T go to infinity together. As a further statistical application, a novel autocovariance test is proposed to detect the equivalence of spiked eigenvalues for two high-dimensional time series. Various simulation stud- ies are illustrated to justify the theoretical findings. Furthermore, a hierarchical clustering approach based on the autocovariance test is constructed and applied to clustering mortality data from multiple countries.

A General m-Estimation Theory in Semi-Supervised Framework

♦*Shanshan Song[1], Yuanyuan Lin[1] and Yong Zhou[2]*
[1]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China
[2]KLATASDS-MOE, School of Statistics and Academy of Statistics andInterdisciplinary Sciences, East China Normal University, Shanghai, China

We study a class of general M-estimators in the semi-supervised setting, wherein the data are typically a combination of a relatively small labeled data set and large amounts of unlabeled data. A new estimator, which efficiently utilises the useful information contained in the unlabeled data, is proposed via a projection technique. We prove consistency and asymptotic normality, and provide an inference procedure based on K-fold cross-validation. The optimal weights are derived to balance the con- tributions of the labelled and unlabelled data. It is shown that the proposed method, by taking advantage of the unlabelled data, produces asymptotically more efficient estimation of the target parameters than the supervised counterpart. Supportive numerical evidence is shown in simulation studies. Applications are illustrated in analysis of the homeless data in Los Angeles.

Semi-Supervised Inference for Nonparametric Logistic Regression

♦*Tong Wang[1], Wenlu Tang[2], Yuanyuan Lin[1] and Wen Su[3]*
[1]Department of Statistics, The Chinese University of Hong Kong
[2]Department of Applied Mathematics, The Hong Kong Polytechnic University
[3]Department of Statistics and Actuarial Science, The University of Hong Kong

We consider the problem of estimating the nonparametric function in nonparametric logistic regression under semi-supervised framework, where a relatively small size labeled data set collected by case-control sampling and a relatively large size of unlabeled data containing only observations of predictors are available. This problem arises in various applications when the outcome variable is expensive or difficult to be observed directly. A two-stage nonparametric semi-supervised estimator based on spline method is proposed to estimate the target regression function by maximizing the likelihood function of the labeled case-control data. The unlabeled data are used in the first stage for estimating the density function that involves in the likelihood function. The consistency and functional asymptotic normality of the semi-supervised two-stage estimator are established under mild conditions. The proposed method, by making use of the unlabeled data, produces more efficient estimation of the target function than the traditional supervised counterpart. The performance of the proposed method is evaluated through extensive simulation studies. An application is

illustrated with an analysis of a skin segmentation data.

Session 23INT54: Statistical Methods in Health Research

Linkage of Big Data of Electronic Medical Records in the Presence of Missing Data

◆*Xiaochun Li*[1], *Huiping Xu*[1] *and Shaun Grannis*[2]
[1]Department of Biostatistics and Health Data Science, School of Medicine, Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana 46202, U.S.A.
[2]Regenstrief Institute, Inc., Indianapolis, IN

Background: Quality patient care requires comprehensive health care data from a broad set of sources. Electronic medical records (EMR) are increasingly distributed across many sources as our nation moves into an era of electronic health record systems. But EMR data are often from independent databases without a common patient identifier, the lack of which impedes data aggregation, causes waste (e.g., tests repeated unnecessarily), affects patient care and hinders research. Record Linkage is the first requite step before effective and efficient patient care and research. Absent a unique universal patient identifier, linkage of patient records is a non-trial task. In addition, the ubiquity of missing data in EMR poses further challenges in record linkage. Objectives: We address the real-world challenges of missing data and matching field selection in linking medical records and evaluate the extent to which incorporating the missing-at-random assumption in the Fellegi-Sunter model and using data-driven selected fields improve patient matching accuracy using real-world use cases. Methods: We incorporated the missing data in the Fellegi-Sunter model using the missing-at-random assumption and compared the proposed approach to the common strategy of treating missing values as disagreement, with matching fields specified by experts or selected by data-driven methods. We used four use cases, each containing a random sample of record pairs with match status ascertained by manual reviews. Use cases included health information exchange (HIE) records deduplication, linkage of public health registry records to HIE, linkage of Social Security Death Master File records to HIE, and newborn screening records deduplication, representative of real-work clinical and public health scenarios. Matching performance was evaluated using sensitivity, specificity, positive predictive value, negative predictive value and F-score. Results: Incorporating the missing-at-random assumption in the Fellegi-Sunter model maintained or improved F-scores whether matching fields were expert-specified or selected by data-driven methods. Combining the missing-at-random assumption and data-driven fields produced the best F-scores in the four use cases. Conclusions: Missing-at-random is a reasonable assumption in real-world record linkage applications: it maintains or improves F-scores regardless of whether matching fields are expert-specified or data-driven. Data-driven selection of fields coupled with MAR achieves the best overall performance, which can be especially useful in privacy-preserving record linkage.

Independent Screening for Nonparametric Additive Cox Model

Jiancheng Jiang

Survival data with ultrahigh dimensional covariates are increasingly common recently due to the rapid development in technologies. It is challenging to model them using survival models in order to understand the association between covariate information and clinical information. In this paper, we focus on the nonparametric additive Cox's proportional model and propose an independent screening method for ultrahigh dimensional data. The proposed screening method is based on the favored bandwidth of the local partial likelihood estimator. Moreover, we develop a two-step procedure to recover all important covariates. This procedure rst captures important variables with nonlinear impacts, and then identi es important variables with linear impacts. We further prove that the nonlinear screening achieves the model selection consistency. Monte Carlo simulations are carried out to evaluate the performance of the proposed screening procedure, which provides evidence supporting the theory. Furthermore, we demonstrate the proposed methodology via a real data example.

Assessing Intra- And Inter-Method Agreement of Functional Data

Ye Yue[1], ◆*Jeong Hoon Jang*[2] *and Amita Manatunga*[1]
[1]Emory University
[2]Yonsei University

With advancement in technology, modern medical devices are increasingly producing complex data that could potentially offer deeper insights into physiological mechanisms underlying diseases and aid in improving diagnostic capabilities. One type of complex data that arises frequently in medical imaging studies is functional data whose sampling unit is a smooth continuous function defined over a time or spatial domain. In this work, we propose a series of intraclass correlation coefficient (ICC) and concordance correlation coefficient (CCC) indices that can evaluate the reliability and reproducibility of medical devices producing functional data. Specifically, two versions of ICC and CCC indices are introduced. The first version consists of time-dependent ICC and CCC indices that can quantify the degrees of intra-method, inter-method and total (intra+inter) agreement that vary smoothly over time. The second version denote their global counterparts that summarize agreement over the entire dime domain using a single measure. The proposed indices are formulated based on a multivariate multilevel functional model that represent indices in terms of truncated multivariate Karhunen-Lo'eve expansions, whose terms can be smoothly estimated by functional principal component analysis. Extensive simulation studies are performed to assess the finite-sample properties of the estimators. The proposed method is applied to Emory renal study data to evaluate the reliability and reproducibility of renogram curve data produced by a high-tech radionuclide image scan that is used to non-invasively detect kidney obstruction.

Assessing Disparities in Americans' Exposure to Pcbs and Pbdes Based on Nhanes Pooled Biomonitoring Data

Yan Liu[1], ◆*Dewei Wang*[2], *Li Li*[1] *and Dingsheng Li*[1]
[1]University of Nevada, Reno
[2]University of South Carolina

The National Health and Nutrition Examination Survey (NHANES) has been continuously biomonitoring Americans' exposure to two families of harmful environmental chemicals: polychlorinated biphenyls (PCBs) and polybrominated diphenyl ethers (PBDEs). However, biomonitoring these chemicals is expensive. To save cost, in 2005, NHANES resorted to pooled biomonitoring, i.e., amalgamating individual specimens to form

a pool and measuring chemical levels from pools. Despite being publicly available, these pooled data gain limited applications in health studies. Among the few studies using these data, racial/age disparities were detected, but there is no control for confounding effects. These disadvantages are due to the complexity of pooled measurements and a dearth of statistical tools. Herein, we developed a regression-based method to unzip pooled measurements, facilitating a comprehensive assessment of disparities in exposure to these chemicals. We found increasing dependence of PCBs on age and income, whereas PBDEs were the highest among adolescents and seniors and were elevated among the low-income population. In addition, Hispanics had the lowest PCBs and PBDEs among all demographic groups after controlling for potential confounders. These findings can guide the development of population-specific interventions to promote environmental justice. Moreover, both chemical levels declined throughout the period, indicating the effectiveness of existing regulatory policies.

Session 23INT63: Recent Advances in Computational Algorithms for Statistical Inference

Sharp Analysis of Em for Learning Mixtures of Pairwise Differences

Abhishek Dhawan, ♦Cheng Mao and Ashwin Pananjady
Georgia Institute of Technology

We consider a symmetric mixture of linear regressions with random samples from the pairwise comparison design, which can be seen as a noisy version of a type of Euclidean distance geometry problem. We analyze the expectation-maximization (EM) algorithm locally around the ground truth and establish that the sequence converges linearly, providing an l_∞-norm guarantee on the estimation error of the iterates. Furthermore, we show that the limit of the EM sequence achieves the sharp rate of estimation in the l_2-norm, matching the information-theoretically optimal constant. We also argue through simulation that convergence from a random initialization is much more delicate in this setting, and does not appear to occur in general. Our results show that the EM algorithm can exhibit several unique behaviors when the covariate distribution is suitably structured.

Estimation of Leading Multi-Block Canonical Correlation Directions via ℓ_1-Norm Constrained Proximal Gradient Descent

Leying Guan
Yale University

Multi-block CCA constructs linear relationships that explain coherent variations across multiple data blocks. We view it as a generalized eigenvalue problem and estimate the leading mCCA direction using ℓ_1-constrained proximal gradient descent. Rather than fixing one constraint value, we propose a decaying sequence of constraints over successive iterations and show that the resulting estimate is rate-optimal under suitable assumptions. While previous work has demonstrated such optimality with a fixed ℓ_0 constraint, the same level of theoretical understanding is still lacking for the ℓ_1 constrained formulation. It is unclear whether there exists a ℓ_1-constrained formulation that can achieve the optimal rate, despite being widely used in practice. In addition, we describe a simple deflation procedure for sequentially estimating multiple directions. We compare our proposal to several existing methods whose implementations are

available on R CRAN, and the proposed method outperforms its competitors in both simulations and a TCGA cancer data set.

Fundamental Limits of Spectral Clustering in Stochastic Block Models

Anderson Ye Zhang
University of Pennsylvania

We give a precise characterization of the performance of spectral clustering for community detection under Stochastic Block Models by carrying out sharp statistical analysis. We show spectral clustering has an exponentially small error with matching upper and lower bounds that have the same exponent, including the sharp leading constant. The fundamental limits established for the spectral clustering hold for networks with multiple and imbalanced communities and sparse networks with degrees far smaller than $\log n$. The key to our results is a novel truncated ℓ_2 perturbation analysis for eigenvectors and a new analysis idea of eigenvectors truncation.

Session 23INTKT3: Keynote Talk 3: Ji Zhu

Statistical Inference on Latent Space Models for Network Data

Ji Zhu
University of Michigan

Recent advances in computing and measurement technologies have led to an explosion in the amount of data with network structures in a variety of fields including social networks, biological networks, transportation networks, the World Wide Web, and so on. This creates a compelling need to understand the generative mechanism of these networks and to explore various characteristics of the network structures in a principled way. Latent space models are powerful statistical tools for modeling and understanding network data. While the importance of accounting for uncertainty in network analysis is well recognized, current literature predominantly focuses on point estimation and prediction, leaving the statistical inference of latent space network models an open question. In this talk, I will present some of our recent work that aims to fill this gap by providing a general framework for analyzing the theoretical properties of the maximum likelihood estimators for latent space network models. In particular, we establish uniform consistency and individual asymptotic distribution results for latent space network models with a broad range of link functions and edge types. Furthermore, the proposed framework enables us to generalize our results to the sparse and dependent-edge scenarios. Our theories are supported by simulation studies and have the potential to be applied in downstream inferences, such as link prediction and network-assisted supervised learning.

Session 23INT101: Statistical Machine Learning and Inference

Simple-Rc: Group Network Inference with Non-Sharp Nulls and Weak Signals

Jianqing Fan[1], Yingying Fan[2], ♦Jinchi Lv[2] and Fan Yang[3]
[1]Princeton University
[2]University of Southern California
[3]Tsinghua University

Large-scale network inference with uncertainty quantification has important applications in natural, social, and medical sciences. The recent work of Fan, Fan, Han and Lv (2022) introduced a general framework of statistical inference on membership profiles in large networks (SIMPLE) for testing the sharp null hypothesis that a pair of given nodes share the same membership profiles. In real applications, there are often groups of nodes under investigation that may share similar membership profiles at the presence of relatively weaker signals than the setting considered in SIMPLE. To address these practical challenges, in this paper we propose a SIMPLE method with random coupling (SIMPLE-RC) for testing the non-sharp null hypothesis that a group of given nodes share similar (not necessarily identical) membership profiles under weaker signals. Utilizing the idea of random coupling, we construct our test as the maximum of the SIMPLE tests for subsampled node pairs from the group. Such technique reduces significantly the correlation among individual SIMPLE tests while largely maintaining the power, enabling delicate analysis on the asymptotic distributions of the SIMPLE-RC test. Our method and theory cover both the cases with and without node degree heterogeneity. These new theoretical developments are empowered by a second-order expansion of spiked eigenvectors under the ℓ_∞-norm, built upon our work for random matrices with weak spikes. Our theoretical results and the practical advantages of the newly suggested method are demonstrated through several simulation and real data examples. This is a joint work with Jianqing Fan, Yingying Fan and Fan Yang.

A Non-Asymptotic Framework for the Approximate Message Passing Algorithm

Yuting Wei
University of Pennsylvania

Approximate message passing (AMP) emerges as an effective iterative paradigm for solving high-dimensional statistical problems. However, prior AMP theory — which focused mostly on high-dimensional asymptotics — fell short of predicting the AMP dynamics when the number of iterations surpasses $o(\log n / \log \log n)$ (with n the problem dimension). To address this inadequacy, in this talk, we introduce a non-asymptotic framework for understanding AMP. Built upon a new decomposition of AMP updates and controllable residual terms, we lay out an analysis recipe to characterize the finite-sample convergence of AMP. As concrete consequences of the proposed analysis recipe: (i) when solving Z2 synchronization, we predict the behavior of randomly initialized AMP for up to $O(n/poly(\log n))$ iterations, showing that the algorithm succeeds without the need of a careful spectral initialization and also a subsequent refinement stage (as conjectured recently by Celentano et al.); (ii) we also characterize the non-asymptotic behavior of AMP in sparse PCA (in the spiked Wigner model) for a broad range of signal-to-noise ratio.

Fairness-Adjusted Neyman-Pearson Classifiers

Ziqing Guo[1], Xin Tong[2] and ♦Lucy Xia[1]
[1]HKUST
[2]USC

Automated algorithmic decision-making is an essential process for many organizations, and developing efficient statistical methods for this purpose is a top priority. However, achieving organizational efficiency is a complex task since multiple aspects need to be optimized simultaneously. These aspects include al-

gorithmic fairness, i.e., minimizing systemic bias against certain disadvantaged social groups, and economic efficiency, i.e., minimizing the cost induced by incorrect decisions. To address these targets, we utilize a dual-focused Neyman-Pearson (NP) classification paradigm that seeks minimal type II error under simultaneous control over both type I error and fairness bias. Leveraging an LDA model, we develop a new oracle framework for dual-focused NP classification, which is a first of its kind. Our proposed finite-sample-based classifiers satisfy both the fairness constraint and type I error constraint with high probability at the population level. We also derive oracle bounds on the excess type II error. Notably, our new classifier does not require sample splitting, which was necessary for most existing NP methods, leading to further increased data efficiency. Numerical and real data analyses demonstrate its superior performance.

Online Tensor Learning: Computational and Statistical Trade-Offs, Adaptivity and Optimal Regret

Jian-Feng Cai, Jingyang Li and ♦Dong Xia
HKUST

We investigate a generalized framework for estimating latent low-rank tensors in an online setting, encompassing both linear and generalized linear models. This framework offers a flexible approach for handling continuous or categorical variables. We propose an online Riemannian gradient descent algorithm, which demonstrates linear convergence and the ability to recover the low-rank component under appropriate conditions. Interestingly, online methods allow us to establish a precise entry-wise error bound for noisy tensor completion. We observe a surprising trade-off between computational and statistical efficiencies. Increasing the step size accelerates convergence but leads to higher statistical error, whereas a smaller step size yields a statistically optimal estimator at the expense of slower convergence. Moreover, we conduct regret analysis for online tensor regression. Under the fixed step size regime, a fascinating trilemma concerning the convergence rate, statistical error rate, and regret is observed. With a proper choice of step size we achieve a square-root regret bound . We extend our analysis to the adaptive setting where the horizon T is unknown. By an adaptive step size schedule , we can attain a statistically optimal error rate along with an optimal logarithmic-order regret bound. To validate our theoretical claims, we provide numerical results that corroborate our findings and support our assertions.

Session 23INT3: New Machine Learning and Semiparametric Methods for Personalized Medical Decision Making

Linear Discriminant Analysis with High-Dimensional Mixed Variables

Binyan Jiang
The Hong Kong Polytechnic University Shenzhen Research Institute

Datasets containing both categorical and continuous variables are frequently encountered in many areas. The dimensions of these variables can be very high especially in modern data analysis. Despite the recent progress made in modelling high-dimensional data for continuous variables, there is a scarcity of methods that can deal with a mixed set of vari- ables. To fill this gap, this paper develops a novel approach for classi-

fying high-dimensional observations with mixed variables. Our framework builds on a location model, in which the distributions of the continuous variables conditional on categorical ones are assumed Gaussian. We overcome the challenge of having to split data into exponentially many cells, or combinations of the categorical variables, by kernel smoothing, and provide new perspectives for its bandwidth choice to ensure an analogue of Bochner's Lemma, which is different to the usual bias-variance tradeoff. We show that the two sets of parameters in our model can be separately estimated and provide penalized likelihood for their estimation. Results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by extensive simulation and real data studies.

Recommending when to Treat: From Binary to Time-to-Intervention Decision

Li Hsu[1], ♦Yair Goldberg[2] and Yingye Zheng[1]
[1]Fred Hutchinson Cancer Research Center
[2]Technion - Israel Institute of Technology

Precision medicine has the potential to improve the practice of disease prevention and treatment. For many complex diseases, factors such as lifestyle factors, environmental factors, and owing to high throughput –omics technologies genetic risk factors have already been identified. This has raised the expectation that the risk prediction models built upon these risk factors can substantially improve prediction accuracy. It is thus important to understand how the model can be used in clinical practice. It is common to use the model to make a binary decision, e.g., whether or not a test should be offered given the subject's risk profile. Many measures have been proposed to evaluate the usefulness of a model with such a binary decision. However, sometimes it is also of interest to know "when to treat". In this talk, I will present a novel concept, "recommended time to start intervention" based on the subject's risk profiles and the time-dependent risk prediction model. I will also present time-dependent measures for assessing the usefulness of the model with the "when-to-treat" decision. This will add to the tools that patients, providers, and policymakers can use to make individualized decisions, which ultimately improve the patients' health without unnecessary treatments or tests.

Matching-Based Learning for Decision Making using Electronic Health Records

Yuanjia Wang
Columbia University

Electronic health records (EHRs) collected from large-scale health systems provide rich subject-specific information on a broad patient population at a lower cost compared to randomized controlled trials. Thus, EHRs may serve as a complementary resource to provide real-world data to construct individualized treatment rules (ITRs) and achieve precision medicine. However, in the absence of randomization, inferring treatment rules from EHR data may suffer from unmeasured confounding. In this article, we propose a self-matched learning method inspired by the self-controlled case series (SCCS) design to mitigate this challenge. We alleviate unmeasured time-invariant confounding between patients by matching different periods of treatments within the same patient (self-controlled matching) to infer the optimal ITRs. The proposed method constructs a within-subject matched value function for optimizing ITRs and bears similarity to the SCCS design. We examine assumptions

that ensure the validity of our method. Sensitivity analyses show that the proposed method is robust under different scenarios. Finally, we apply self-matched learning to estimate the optimal ITRs from type 2 diabetes patient EHRs, which shows our estimated decision rules lead to greater advantages in reducing patients' diabetes-related complications.

Learning Optimal Group-Structured Individualized Treatment Rules

Haixu Ma, ♦Donglin Zeng and Yufeng Liu
University of North Carolina

Personalized medicine aims to determine the optimal Individualized Treatment Rule (ITR) tailored to each patient's characteristics. When many treatment options are present, existing methods suffer from data insufficiency to estimate the optimal ITR. On the other hand, it is often observed that some treatments have similar individual effects due to the same mechanism of drug action. Therefore, we propose GRoup Outcome Weighted Learning (GROWL) to learn the latent group structure among treatments, where treatments within the same group have no difference when treating individual patients, and at the same time, to estimate the optimal group-structured ITRs through a single optimization. Fisher consistency, the excess risk bound, and the convergence rate of the value function are established to provide a theoretical guarantee for GROWL. Extensive empirical results in simulation studies and real data analysis demonstrate that GROWL has better performance than other existing methods.

Session 23INT45: Recent Developments in Statistical Network Analysis

Higher-Order Accurate Two-Sample Network Inference and Network Hashing

Meijia Shao[1], Dong Xia[2], ♦Yuan Zhang[1], Qiong Wu[3] and Shuo Chen[4]
[1]The Ohio State University
[2]Hong Kong University of Science and Technology
[3]University of Pennsylvania
[4]University of Maryland, Baltimore

Two-sample hypothesis testing for comparing two networks is an important yet difficult problem. Major challenges include: potentially different sizes and sparsity levels; non-repeated observations of adjacency matrices; computational scalability; and theoretical investigations, especially on finite-sample accuracy and minimax optimality. In this article, we propose the first provably higher-order accurate two-sample inference method by comparing network moments. Our method extends the classical two-sample t-test to the network setting. We make weak modeling assumptions and can effectively handle networks of different sizes and sparsity levels. We establish strong finite-sample theoretical guarantees, including rate-optimality properties. Our method is easy to implement and computes fast. We also devise a novel nonparametric framework of offline hashing and fast querying particularly effective for maintaining and querying very large network databases. We demonstrate the effectiveness of our method by comprehensive simulations. We apply our method to two real-world data sets and discover interesting novel structures.

Limit Results for Distributed Estimation of Invariant Subspaces in Multiple Networks Inference and Pca

Runbing Zheng and ♦Minh Tang
North Carolina State University

We study the problem of estimating the left and right singular subspaces for a collection of heterogeneous random graphs with a shared common structure. We analyze an algorithm that first estimates the orthogonal projection matrices corresponding to these subspaces for each individual graph, then computes the average of the projection matrices, and finally finds the matrices whose columns are the eigenvectors corresponding to the d largest eigenvalues of the sample averages. We show that the algorithm yields an estimate of the left and right singular vectors whose row-wise fluctuations are normally distributed around the rows of the true singular vectors. We then consider a two-sample hypothesis test for the null hypothesis that two graphs have the same edge probabilities matrices against the alternative hypothesis that their edge probabilities matrices are different. Using the limiting distributions for the singular subspaces, we present a test statistic whose limiting distribution converges to a central 2 (resp. non-central 2) under the null (resp. alternative) hypothesis. Finally, we adapt the theoretical analysis for multiple networks to the setting of distributed PCA; in particular, we derive normal approximations for the rows of the estimated eigenvectors using distributed PCA when the data exhibit a spiked covariance matrix structure.

Subsampling-Based Modified Bayesian Information Criterion for Large-Scale Stochastic Block Models

Jiayi Deng[1], ♦Danyang Huang[1], Xiangyu Chang[2] and Bo Zhang[1]
[1]Renmin University of China
[2]Xi'an Jiaotong University

Identifying the number of communities is a fundamental problem in community detection, which has received increasing attention recently. However, rapid advances in technology have led to the emergence of large-scale networks in various disciplines, thereby making existing methods computationally infeasible. To address this challenge, we propose a novel subsampling-based modified Bayesian information criterion (SM-BIC) for identifying the number of communities in a network generated via the stochastic block model and degree-corrected stochastic block model. We first propose a node-pair subsampling method to extract an informative subnetwork from the entire network, and then we derive a purely data-driven criterion to identify the number of communities for the subnetwork. In this way, the SM-BIC can identify the number of communities based on the subsampled network instead of the entire dataset. This leads to important computational advantages over existing methods. We theoretically investigate the computational complexity and identification consistency of the SM-BIC. Furthermore, the advantages of the SM-BIC are demonstrated by extensive numerical studies.

Learning Network Properties without Network Data – a Correlated Network Scale-Up Model

Ian Laga[1], Le Bao[2] and ♦Xiaoyue Niu[2]
[1]Montana State University
[2]Penn State University

The network scale-up method based on "how many X's do you know?" questions has gained popularity in estimating the sizes of hard-to-reach populations. The success of the method re-lies primarily on the easy nature of the data collection and the flexibility of the procedure, especially since the model does not require a sample from the target population, a major limitation of traditional size estimation models. In this talk, we propose a new network scale-up model which incorporates respondent and subpopulation covariates in a regression framework and includes a bias term that is correlated between subpopulations. We also introduce a new scaling procedure utilizing the correlation structure. In addition to estimating the unknown population sizes, our proposed model depicts people's social network patterns in an aggregated level without using the network data.

Session 23INT43: Modern Machine Learning Approaches for Efficient Estimation and Sampling

Bayesian Analysis for Functional Anova Model

Yongdai Kim
Seoul National University

Functional ANOVA model is a useful tool to construct an interpretable prediction model. While there are several frequentist procedures to estimate the components in the functional ANOVA model (i.e. MARS, Splines), Bayesian procedures focus mainly on the generalized additive model (GAM) that is the simplest one among functional ANOVA model due to computational difficulties. In this talk. we propose a computationally efficient Bayesian procedure to infer components in the functional ANOVA model. Our algorithm, called ANOVA-BART, is a modification of BART (Bayesian Additive Regression Tree), a well knonw Bayesian procedure for estimating high-dimensional regression model. BART combines many baseline trees (simple tress) to infer the final prediction model. Even though BART is very good at prediction, its interpretation is no easy. To improve interpretability of BART, we contruct the sets of baseline trees corresponding to each component of the functional ANOVA model, put prior on each set of baseline trees, and develop an MCMC algorithm to search good linear combinations of baseline trees for each component.

Robust Estimation of Central Subspace under High-Dimensional and Elliptically-Contoured Design

♦Jing Zeng[1] and Qing Mai[2]
[1]University of Science and Technology of China
[2]Florida State University

Sufficient dimension reduction (SDR) is a valuable tool to tackle high dimensionality while maintaining the primary information of the prediction problem, and it has demonstrated great promise in many applications. There exist a variety of high-dimensional sufficient dimension reduction methods in the literature. However, they all rely on the sub-Gaussian assumption of the predictors' marginal distribution or conditional distribution. Such a light-tailedness assumption is frequently violated in real life. A new methodology is proposed to estimate the central subspace consistently when the predictor exhibits heavy-tailedness. Our novel proposal overcomes both the heavy-tailedness and the high dimensionality. Under a general regression model assumption and the elliptically-contoured design assumption, an invariance result between the CS and a surrogate subspace is established. Estimating the surrogate subspace avoids the heavy-tailedness issue and can be implemented using existing high-dimensional SDR methods. Theoretically,

our proposal enjoys satisfactory consistency, and the convergence rate is shown to achieve optimality. Empirically, the efficiency and effectiveness of our recommendation are demonstrated by extensive simulation studies and real data examples.

Data-Adaptive Discriminative Feature Localization with Statistically Guaranteed Interpretation

◆Ben Dai[1], Xiaotong Shen[2], Lin Yee Chen[2], Chunlin Li[2] and Wei Pan[2]

[1]The Chinese University of Hong Kong

[2]University of Minnesota

In explainable artificial intelligence, discriminative feature localization is critical to reveal a blackbox model's decision-making process from raw data to prediction. In this article, we use two real datasets, the MNIST handwritten digits and MIT-BIH Electrocardiogram (ECG) signals, to motivate key characteristics of discriminative features, namely adaptiveness, predictive importance and effectiveness. Then, we develop a localization framework based on adversarial attacks to effectively localize discriminative features. In contrast to existing heuristic methods, we also provide a statistically guaranteed interpretability of the localized features by measuring a generalized partial R2. We apply the proposed method to the MNIST dataset and the MIT-BIH dataset with a convolutional auto-encoder. In the first, the compact image regions localized by the proposed method are visually appealing. Similarly, in the second, the identified ECG features are biologically plausible and consistent with cardiac electrophysiological principles while locating subtle anomalies in a QRS complex that may not be discernible by the naked eye. Overall, the proposed method compares favorably with state-of-the-art competitors. Accompanying this paper is a Python library dnn-locate (this https URL) that implements the proposed approach.

Efficient Multimodal Sampling via Tempered Distribution Flow

◆Yixuan Qiu[1] and Xiao Wang[2]

[1]Shanghai University of Finance and Economics

[2]Purdue University

Sampling from high-dimensional distributions is a fundamental problem in statistical research and practice. However, great challenges emerge when the target density function is unnormalized and contains isolated modes. We tackle this difficulty by fitting an invertible transformation mapping, called a transport map, between a reference probability measure and the target distribution, so that sampling from the target distribution can be achieved by pushing forward a reference sample through the transport map. We theoretically analyze the limitations of existing transport-based sampling methods using the Wasserstein gradient flow theory, and propose a new method called TemperFlow that addresses the multimodality issue. TemperFlow adaptively learns a sequence of tempered distributions to progressively approach the target distribution, and we prove that it overcomes the limitations of existing methods. Various experiments demonstrate the superior performance of this novel sampler compared to traditional methods, and we show its applications in modern deep learning tasks such as image generation.

Session 23INT25: Recent Development of Statistical Methods for Health Sciences

Two-Sample Test and Support Recovery for Image Data

◆Lianqiang Qu[1], Jian Huang, Liuquan Sun and Hongtu Zhu
[1]Central China Normal University

We propose a multiscale adaptive test for detecting differences between two samples of image data, which are intrinsically smoothed. The proposed test aggregates data from nearby locations, dramatically improving statistical power. A salient feature of our proposed test is its adaptivity to the spatial features of image data. We show that the proposed test statistic converges to a type I extreme value distribution under the null hypothesis. Furthermore, we investigate the multiscale nature of the proposed test and show that the chosen scales can grow at a certain polynomial rate of the sample size. We also evaluate its power against sparse alternatives and demonstrate that when the null hypothesis of equal means is rejected, the multiscale adaptive test can identify the locations where the two samples differ from each other with probability tending to one. Additionally, we extend the proposed method to multi-sample ANOVA tests. Simulation results suggest that the proposed tests outperform naive methods that do not consider the spatial features of imaging data. The procedures are illustrated on a real dataset from the Alzheimer's Disease Neuroimaging Initiative study.

Deep Learning for Time-to-Event Predictions with Applications to Ehr Data

Xueying Wang[1], Jing Ning[2], Ruosha Li[1], Han Feng[3] and ◆Hulin Wu[1]
[1]University of Texas Health Science Center at Houston
[2]University of Texas MD Anderson Cancer Center
[3]Tulane University

With the rapid development of electronic health records (EHR) systems, the large and diverse clinical information in EHR systems has become an important data source for clinical and epidemiological research. It is interesting to predict the mortality risk after diagnosis of a disease or clinical condition within a pre-specified time t, known as t-year risk predictions. Recently, machine learning approaches have been proposed to deal with complex predictors for time-to-event outcomes subject to censoring. However, most methods do not appropriately account for censoring in the machine learning literature. In this study, we propose a bias correction method for risk predictions of an event with censored data by incorporating the inverse probability of censoring weights (IPCW) in predictive machine learning models for dichotomous outcomes. We apply the IPCW-weighted bias-correction machine learning and survival model-based methods to predict the t-year risk of heart failure (HF) after diagnosis of type 2 diabetes mellitus (T2DM) based on a nationwide large EHR database. The EHR data application illustrates a good example that the IPCW-weighted bias-corrected machine learning models outperform the survival model-based methods for t-year risk predictions in a real-world dataset. A simulation study is conducted to demonstrate the necessity of bias correction. This study is part of a PhD student, Xueying Wang's dissertation by collaboration with Drs. Jing Ning, Ruosha Li and Han Feng.

Theoretical Properties of Oversampling and Subsampling for Imbalanced Classification

Jie Zhou

TBC

Post-Episodic Reinforcement Learning Inference

Vasilis Syrgkanis[1] *and* ♦*Ruohan Zhan*[2]
[1]Stanford University
[2]Hong Kong University of Science and Technology

We consider estimation and inference with data collected from episodic reinforcement learning (RL) algorithms; i.e. adaptive experimentation algorithms that at each period (aka episode) interact multiple times in a sequential manner with a single treated unit. Our goal is to be able to evaluate counterfactual adaptive policies after data collection and to estimate structural parameters such as dynamic treatment effects, which can be used for credit assignment (e.g. what was the effect of the first period action on the final outcome). Such parameters of interest can be framed as solutions to moment equations, but not minimizers of a population loss function, leading to Z-estimation approaches in the case of static data. However, such estimators fail to be asymptotically normal in the case of adaptive data collection. We propose a re-weighted Z-estimation approach with carefully designed adaptive weights to stabilize the episode-varying estimation variance, which results from the nonstationary policy that typical episodic RL algorithms invoke. We identify proper weighting schemes to restore the consistency and asymptotic normality of the re-weighted Z-estimators for target parameters, which allows for hypothesis testing and constructing uniform confidence regions for target parameters of interest. Primary applications include dynamic treatment effect estimation and dynamic off-policy evaluation. This is joint work with Vasilis Syrgkanis.

Session 23INT70: Quantile Regression with Complex Data

A Semiparametric Quantile Single-Index Model for Zero-Inflated and Overdispersed Outcomes

Tianying Wang
Tsinghua University

We consider the complex data modeling problem motivated by the zero-inflated and overdispersed microbiome read count data. Several parametric approaches have been proposed to address issues of zero inflation and overdispersion, such as zero-inflated Poisson regression and zero-inflated Negative Binomial regression. However, parametric assumptions could be easily violated in real-world applications. To relax the parametric assumptions and provide a robust modeling framework, we consider single-index quantile regression models, as quantile regression makes no distribution assumptions, and single-index models provide more flexibility than linear models while remaining interpretability. Though single-index models have been studied under the quantile regression framework, they cannot be applied directly to zero-inflated outcomes, especially when the degree of zeros varies across subjects. We propose a semiparametric single-index quantile regression framework, which is flexible to include a wide range of possible association functions and adaptable to the various zero proportions across subjects. We establish the asymptotic normality of the index coefficients estimator and the asymptotic convergence rate of the nonparametric quantile regression curve estimation. Through Monte Carlo simulation studies and the application in a microbiome study, we demonstrate the superior performance of the proposed method.

Fast Imputation Algorithms in Quantile Regression with Missing Covariates

Hao Cheng
National Academy of Innovation Strategy, China Association for Science and Technology

In many applications, some covariates could be missing for various reasons. Regression quantiles could be either biased or under-powered when ignoring the missing data. Multiple imputation and EM-based augment approach have been proposed to fully utilize the data with missing covariates for quantile regression. Both methods however are computationally expensive. We propose a fast imputation algorithm (FI) to handle the missing covariates in quantile regression, which is an extension of the fractional imputation in likelihood based regressions. FI and modified imputation algorithms (FIIPW and MIIPW) are compared to existing MI and IPW approaches in the simulation studies, and applied to two real data examples.

Censored Quantile Regression Forest

♦*Huichen Zhu*[1], *Yifei Sun*[2] *and Ying Wei*[2]
[1]The Chinses University of Hong Kong
[2]Columbia University

In various scenarios, determining the impact of different treatments on a censored response variable is crucial, and it is natural to assess these effects at different quantiles (e.g., median). However, the presence of right censoring, the unknown structure of treatment effects, and the large number of potential effect modifiers present significant challenges. To address these issues, we propose a forest approach, which employs a combination of random forests and censored quantile regression to assess heterogeneous effects that vary with high-dimensional variables. Additionally, we propose a variable importance decomposition to measure the impact of a variable on the treatment effect function.

Session 23INT1: High-Dimensional Data Analysis

Ensemble Projection Pursuit for General Nonparametric Regression

Haoran Zhan, Mingke Zhang and ♦*Yingcun Xia*

Projection Pursuit Regression (PPR) has played an important role in the development of statistics and machine learning. However, as a statistical learning method, PPR has not yet demonstrated an accuracy comparable to other methods such as Random Forests (RF) and Artificial Neural Networks (ANN). In this paper, we revisit the estimation of PPR and propose a greedy algorithm and an ensemble approach via feature bagging, hereafter referred to as ePPR. Compared to Random Forest (RF), ePPR has two main advantages: (1) its theoretical consistency can be proved for more general regression functions, as long as they are continuous, and higher consistency rates can be obtained; and (2) ePPR does not split the samples, so each term of the PPR is estimated using the whole data, which makes the estimation more efficient and guarantees the smoothness of the estimator. ePPR is also easier to tune and train than ANN. Extensive comparisons on real data sets show that ePPR is noticeably more efficient in regression and classification than RF and other competitors.

Inference on High-Dimensional Single-Index Models with Streaming Data

Dongxiao Han

Nankai University

t is well-known that streaming data brings new challenges to traditional statistical methods. The main difficulties are that data grows rapidly in volume and velocity, and it is infeasible to store such huge entire datasets in memory. In this paper, we present an online inference framework for regression parameters in high-dimensional semiparametric single-index models with an unknown link function. The proposed on- line procedure is updated with only the current data batch and summary statistics of historical data, which completely bypasses re-accessing the raw entire data. Mean- while, we do not require estimating the unknown link function, which can be highly challenging. In addition, the proposed inference procedure is developed based on gen- eral convex loss functions. The Huber loss function and the negative log-likelihood of logistic regression model are provided as examples to illustrate the proposed method. The ℓ_1 and ℓ_2 bounds of the proposed online Lasso estimators and the asymptotic normality of the proposed online debiased Lasso estimators are developed. Extensive simulation studies are conducted to evaluate the performance of the proposed method. Applications to the Nasdaq stock and financial distress datasets are provided.

High-Dimensional Covariance Matrices under Dynamic Volatility Models: Asymptotics and Shrinkage Estimation
Yi Ding[1] and ♦*Xinghua Zheng*[2]
[1]University of Macau
[2]Hong Kong University of Science and Technology
We study the estimation of high-dimensional covariance matrix and its eigenvalues under dynamic volatility models. Data under such models have nonlinear dependency both cross-sectionally and temporally. We first investigate the empirical spectral distribution (ESD) of the sample covariance matrix under scalar BEKK models and establish conditions under which the limiting spectral distribution (LSD) is either the same as or different from the i.i.d. case. We then propose a time-variation adjusted (TV-adj) sample covariance matrix and prove that its LSD follows the same Marcenko-Pastur law as the i.i.d. case. Based on the asymptotics of the TV-adj sample covariance matrix, we develop a consistent population spectrum estimator and an asymptotically optimal nonlinear shrinkage estimator of the unconditional covariance matrix.

Inference for Nonstationary Time Series with Varying Periodicity, a Smooth Trend and Covariate Effects
♦*Ming-Yen Cheng*[1], *Shouxia Wang*[1] and *Lucy Xia*[2]
[1]Hong Kong Baptist University
[2]Hong Kong University of Science and Technology
Traditional analysis of a periodic time series assumes its pattern remains the same over the entire time range. However, using ad hoc methods some recent empirical studies in climatology and other fields find the amplitude may change along with time and that has important implications. We develop a formal procedure to detect and estimate change-points in the periodic pattern. Often there is also a smooth trend, and sometimes the period is unknown and there can be other covariate effects. Based on a new model that takes into account all these, a three-step estimation procedure is proposed to estimate accurately the unknown period, change-points and varying amplitude in the periodic component, the trend and the covariate effects. First, we adopt penalized segmented least squares estimation for the unknown period with the trend and covariate effects approximated by B-splines. Then, given the period estimate, we construct a

novel test statistic and use it in binary segmentation to estimate change-points in the periodic component. Finally, given the period and change-point estimates, we estimate the whole periodic component, trend and covariate effects using B-splines. Asymptotic results for the proposed estimators are derived, including consistency of the period and change-point estimators, and asymptotic normality of the estimated periodic sequence, trend and covariate effects. Simulation results demonstrate appealing performance of the new method, and empirical studies show its advantages.

Session 23INT74: Modern Statistical and Machine Learning Modeling of Big Data

Directly Deriving Parameters from Sdss Photometric Images
♦*Fan Wu and Yude Bu*
Stellar atmospheric parameters (effective temperature, surface gravity and metallicity) are funda mental for understanding the formation and evolution of stars and galaxies. Photometric data can provide a low-cost way to estimate these parameters, but traditional methods based on photometric magnitudes have many limitations. In this paper, we propose a novel model called Bayesian Convit, which combines an approximate Bayesian framework with a deep learning method, namely Convit, to derive stellar atmospheric parameters from Sloan Digital Sky Survey (SDSS) images of stars and effectively provide corresponding confidence levels for all the predictions. We achieve high accuracy for Teff and [Fe/H], with $\sigma(\text{Teff}) = 172.37$ K and $([\text{Fe/H}]) = 0.23$ dex. For log g, which is more challenging to estimate from image data, we propose a two-stage approach: (1) classify stars into two categories based on their log g values (> 4 dex or < 4 dex) and (2) regress separately these two subsets. We improve the estimation accuracy of stars with log g > 4 dex significantly to $\sigma(\log g > 4) = 0.052$ dex, which are comparable to those based on spectral data. The final joint result is $\sigma(\log g) = 0.41$ dex. Our method can be applied to large photometric surveys like Chinese Space Station Telescope(CSST) and Large Synoptic Survey Telescope(LSST)

Barycenter Estimation of Positive Semi-Definite Matrices with Bures-Wasserstein Distance
♦*Jingyi Zheng*[1], *Huajun Huang*[1], *Yuyan Yi*[1], *Yuexin Li*[1] and *Shu-Chin Lin*[2]
[1]Auburn University
[2]National Health Research Institutes, Taiwan
Brain-computer interface (BCI) builds a bridge between human brain and external devices by recording brain signals and translating them into commands for devices to perform the user's imagined action. The core of the BCI system is the classifier that labels the input signals as the user's imagined action. The classifiers that directly classify covariance matrices using Riemannian geometry are widely used not only in BCI domain but also in a variety of fields including neuroscience, remote sensing, biomedical imaging, etc. However, the existing Affine-Invariant Riemannian-based methods treat covariance matrices as positive definite while they are indeed positive semi-definite especially for high dimensional data. Besides, the Affine-Invariant Riemannian-based barycenter estimation algorithms become time consuming, not robust, and have convergence issues when the dimension and number of covariance ma-

trices become large. To address these challenges, in this paper, we establish the mathematical foundation for Bures-Wasserstein distance and propose new algorithms to estimate the barycenter of positive semi-definite matrices efficiently and robustly. Both theoretical and computational aspects of Bures-Wasserstein distance and barycenter estimation algorithms are discussed. With extensive simulations, we comprehensively investigate the accuracy, efficiency, and robustness of the barycenter estimation algorithms coupled with Bures-Wasserstein distance. The results show that Bures-Wasserstein based barycenter estimation algorithms are more efficient and robust.

Simultaneous Identification of Brain Functional Differential Network and Gene Regulatory Pathways

Hao Chen[1], Yong He[2] and ♦Jiadong Ji[2]
[1]Peking University
[2]Shandong University

Prior studies have identified ApoE gene is associated with alterations in brain functional connectivity, and Alzheimer's disease (AD) process is primarily related to the abnormal brain functional connectivity. Motivated by these results, we are interested in studying how brain functional connectivity mediates the effects of the ApoE gene on AD. In this article, we propose a valid mediation analysis procedure SDNPI for identifying critical brain functional connectivity that might mediate the effect of the ApoE gene on the AD outcome. The proposed approach includes two steps, first conducting the estimation of individual-specific brain functional connectivity and then constructing the penalized mediation model for binary outcomes. Simulation studies are conducted to assess the performance of the proposed method. We apply the proposed procedure to the ADNI dataset. The identified brain functional connectivity alterations in Early AD versus NC and critical brain regions are consistent with prior experimental studies, emphasizing the practical applicability of our method.

Fine-Mapping Causal Variants using Summary Gwas Statistics with Heritability-Induced Dirichlet Decomposition Prior

Xiang Li and ♦Yan Dora Zhang
The University of Hong Kong

The goal of statistical fine-mapping is to identify causal genetic variants associated with complex traits or diseases. Existing methods usually use discrete mixture priors with a pre-specified maximum number of causal variants, and are likely to get trapped into suboptimal solutions. In this work, we will introduce a novel fine-mapping method using summary GWAS statistics based on continuous global-and-local shrinkage prior named h2-D2. We will also introduce a new method to construct credible set of variants in the framework of continuous priors. Simulation studies demonstrate that h2-D2 outperforms other state-of-art fine-mapping methods including SuSiE and FINEMAP. We will also talk about the application of h2-D2 in prostate cancer data analysis.

Session 23INT75: Recent Advances in Integrative Analysis of Multi-Omics Data

Leveraging Genomic Regulations to Inform Genetic Association Analysis with Complex Traits

Li Hsu
TBC

Discrete Representation Learning for Single-Cell Multi-Omics Data

Xuejian Cui[1], Shengquan Chen[2] and ♦Rui Jiang[1]
[1]Tsinghua University
[2]Nankai University

Single-cell chromatin accessibility sequencing (scCAS) data has been growing continuously at an unprecedented pace, but the inherent high dimensionality and sparsity of scCAS data pose significant challenges to downstream analysis. Although deep learning models, especially variational autoencoder(VAE)-based models, have been widely used to generate feature embedding, the intrinsic Gaussian assumption somewhat contradict with real data, making these models lack interpretability. Here we propose CASTLE, a generative deep learning approach based on the framework of Vector Quantization Variational AutoEncoder (VQVAE) to extract discrete latent features that interpretably characterize scCAS data in an unsupervised manner. We validate the superior performance and robustness of CASTLE for accurate clustering and reasonable visualization compared with the state-of-the-art methods, and we demonstrate the advantages of CASTLE for effective incorporation of reference dataset with various quality.

Bayesian Inference for Non-Invasive Preimplantation Genetic Testing

♦Hao Ge, Ruiqi Zhang, Lei Huang and Xiaoliang Xie
Peking University

In vitro fertilization (IVF) requires pre-implantation genetic testing (PGT) to determine the health of embryos before they are transferred to the uterus. Traditional cell sampling methods are invasive and have increased the rate of miscarriage in pregnant women and decreased the live birth rate of embryos. Therefore, in recent years, many laboratories have developed non-invasive cell sampling methods. Although non-invasive methods are safer, they yield less DNA and lower data quality, which may interfere with the determination of embryo genotype and chromosomal ploidy. We proposed a new Bayesian model that combines the characteristics of non-invasive samples for data analysis of non-invasive PGT. We tested this model on clinical data from many families and found that it had a high accuracy rate.

Statistical Methods for Mediation Analysis with High-Dimensional Omics Mediators

♦Peng Wei[1], Sunyi Chi[1], Zhichao Xu[1], Tianzhong Yang[2], Chunlin Li[2] and Xuelin Huang[1]
[1]The University of Texas MD Anderson Cancer Center
[2]University of Minnesota

Environmental exposures can regulate intermediate molecular phenotypes, such as the transcriptome, metabolome and microbiome, by various mechanisms and thereby lead to different health outcomes. It is of significant scientific interest to unravel the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposure and health traits. Mediation analysis is an important tool for investigating such relationships. However, there are many unique challenges facing high-dimensional mediation analysis with these emerging "omics" mediators. To this end, we extended an R-squared (R2) total mediation effect size measure for continuous outcomes, originally proposed in the single-mediator setting, to the moderate- and high-dimensional mediator settings in the mixed model framework. I will introduce R2-based mediation

analysis with high-dimensional omics mediators for continuous outcomes, time-to-event outcomes, and speeding up confidence interval estimation based on asymptotic results.

Session 23INTSP3: Junior Researcher Award Session

Multi-State Model and Structural Selection for the Analysis of Depressive Symptom Dynamics in Middle-Aged and Older Adults
Chuoxin Ma
BNU-HKBU United International College
Depressive symptoms are increasingly common in middle-aged and older adults and have become a major public health problem. People may experience transitions across different underlying states due to symptom severity fluctuation over a course of many years. Characterizing the dynamics of depression and identifying important risk factors associated with different depressive states may contribute to the understanding of disease mechanism and the development of effective interventions. However, existing research rarely consider the temporal dynamics in the associations between risk factors and different depression states. To fill this gap, we analyze the mental health data from Chinese residents aged 45 and older, and investigate whether the influence of risk factors change over time based on the data from the China Health and Retirement Longitudinal Study (CHARLS). Five types of depressive episode states are defined and the risks of transitions among states are modelled by varying-coefficient multi-state models. Risk factors with time-varying effects, time-independent effects, and null effects are selected via model structure selection method.

Partial Quantile Tensor Regression
Dayu Sun
Indiana University
Tensor data, characterized as multidimensional arrays, have become increasingly prevalent in biomedical research. To handle a tensor predictor in the regression setting, most existing methods concern its effect on the mean response, thereby failing to address the practical interest regarding the predictor s effect on non-average or unusual outcomes. In this work, we propose a partial quantile tensor regression (PQTR) framework, which novel applies the core principle of the partial least squares technique to achieve effective dimension reduction for quantile regression with a tensor predictor. The proposed PQTR algorithm is computationally efficient and scalable to a large size tensor predictor. Moreover, we uncover an appealing latent variable model representation for the new PQTR algorithm, justifying a simple population interpretation of the resulting estimator. We further investigate the connection of the PQTR procedure with an envelope quantile tensor regression (EQTR) model, which defines a general set of sparsity conditions tailored to quantile tensor regression. We prove the root-n consistency of the PQTR estimator under the EQTR model, and demonstrate its superior finite-sample performance compared to benchmark methods through simulation studies. We demonstrate the practical utility of the proposed method via an application to a neuroimaging study of posttraumatic stress disorder (PTSD).

Desiderata for Representation Learning: a Causal Perspective
Yixin Wang
University of Michigan

Representation learning constructs low-dimensional representations to summarize essential features of high-dimensional data. This learning problem is often approached by describing various desiderata associated with learned representations; e.g., that they be non-spurious, efficient, or disentangled. It can be challenging, however, to turn these intuitive desiderata into formal criteria that can be measured and enhanced based on observed data. In this paper, we take a causal perspective on representation learning, formalizing non-spuriousness and efficiency (in supervised representation learning) and disentanglement (in unsupervised representation learning) using counterfactual quantities and observable consequences of causal assertions. This yields computable metrics that can be used to assess the degree to which representations satisfy the desiderata of interest and learn non-spurious and disentangled representations from single observational datasets.

Learning Network-Structured Dependence from Non-Stationary Multivariate Point Process Data
Muhong Gao
Chinese Academy of Science
Learning the sparse network-structured dependence among nodes from multivariate point process data Tii?V has wide applications in information transmission, social science and computational neuroscience. This paper develops new continuous-time stochastic models of the conditional intensity processes ?i(t): t? 0i?V for learning the network structure, underlying an array of non-stationary multivariate counting processes N(t): t? 0for Tii?V . The stochastic mechanism of the model is central to statistical inference of graph parameters relevant to structure recovery, but does not meet key assumptions underlying the commonly used Poisson process, Hawkes process, queuing model and the piecewise deterministic Markov process. This inspires us to introduce a new marked point process for intensity discontinuities, derive the compact representations of their conditional distributions and show the cyclicity property of N(t) driven by recurrence time points. These new theoretical properties further enable us to establish statistical consistency and convergence properties of the proposed penalized M-estimators for graph parameters under mild regularity conditions. Simulation evaluations demonstrate computational simplicity of the proposed method, and increased estimation accuracy over existing methods. Real multiple neuron spike train recordings are analyzed to infer connectivity in neuronal networks.

Session 23INT55: Recent Advances of High-Dimensional Models and Time Series Models

Optimal and Safe Estimation for High-Dimensional Semi-Supervised Learning
Yang Ning
Cornell University
There are many scenarios such as the electronic health records where the outcome is much more difficult to collect than the covariates. In this paper, we consider the linear regression problem with such a data structure under the high dimensionality. Our goal is to investigate when and how the unlabeled data can be exploited to improve the estimation and inference of the regression parameters in linear models, especially in light of the fact that such linear models may be misspecified in data analysis. In particular, we address the following two important

questions. (1) Can we use the labeled data as well as the unlabeled data to construct a semi-supervised estimator such that its convergence rate is faster than the supervised estimators? (2) Can we construct confidence intervals or hypothesis tests that are guaranteed to be more efficient or powerful than the supervised estimators?

Cross-Covariance-Based Estimator of Autocorrelation Operator for Arh(1) Process with Contamination

♦Zhao Chen[1], Liang Hu[2] and Yuwei Zhao[1]
[1]Fudan University
[2]Chinese Academy of Sciences

In recent years, the theory and application of functional time series have become a hot topic. Statistical inference of autoregressive models on Hilbert space (ARH) is a challenging issue. In functional time series, it is usually assumed that the functional data can be observed on a grid of points, but many factors may cause the observation data to be contaminated, such as sampling methods, exogenous factors, linear truncation which may affect the performance of statistical inference. To address the issue of data contamination in practice, a set of first-order ARH models contaminated by white noise on Hilbert space is considered. Inspired by the fact that the autocovariance operator is less affected by contamination, we use the autocovariance operator to construct an estimator of the autoregressive coefficient operator. We first prove a series of theoretical properties of high-order autocovariance operators, and then propose an estimator of the autoregressive coefficient operator based on autocovariance operator for ARH(1) processes. We derive the consistency and asymptotic normality of this estimator in both finite and infinite-dimensional cases. Numerical results show that the proposed method reduces the influence of contamination on functional time series and high-order autocovariance operators, and the autoregressive coefficient operator estimated from it still has good statistical properties under contamination.

Inference for High-Dimensional Linear Models with Locally Stationary Error Processes

Jiaqi Xia, Yu Chen and ♦Xiao Guo
University of Science and Technology of China

Linear regression models with stationary errors are well studied but the non-stationary assumption is more realistic in practice. An estimation and inference procedure for high-dimensional linear regression models with locally stationary error processes is developed. Combined with a proper estimator for the autocovariance matrix of the non-stationary error, the desparsified lasso estimator is adopted for the statistical inference of the regression coefficients under the fixed design setting. The consistency and asymptotic normality of the desparsified estimators is established under certain regularity conditions. Element-wise confidence intervals for regression coefficients are constructed. The finite sample performance of our method is assessed by simulation and real data analysis.

Normal Mixture Quasi-Maximum Likelihood Estimation of Double Autoregressive Models

Zhao Chen[1], Min Chen[2] and ♦Christina Dan Wang[3]
[1]Fudan University
[2]AMSS
[3]NYU Shanghai

Abstract:This paper investigates the estimation of double autoregressive model (DAR) with skewed and heavy-tailed innovation. A new estimation method, the normal mixture quasi-maximum likelihood estimation (NM-QMLE) is proposed to estimate the DAR model with the non Gaussian behavior. Under regularity conditions, the consistency and asymptotic normality are established for NMQMLE. The numerical simulation for the DAR model with heavy-tailed and skewed innovation indicates that the NMQMLE outperforms several commonly adopted QMLE's. Finally, An empirical example on the S& P 500 index illustrates the new findings.

Integrative Analysis of Gaussian Graphical Models

Shuangge Ma
Yale University

Heterogeneity is a hallmark of many complex diseases. This study has been motivated by the unsupervised heterogeneity analysis for complex diseases based on molecular and imaging data, for which, network-based analysis can be more informative than that limited to mean, variance, and other simple distributional properties. In the literature, there has been limited research on network-based heterogeneity analysis, and a common limitation shared by the existing techniques is that the number of subgroups needs to be specified a priori or in an ad hoc manner. We develop a novel approach for heterogeneity analysis based on the Gaussian graphical model. It applies penalization to the mean and precision matrix parameters to generate regularized and interpretable estimates. A fusion penalty is imposed to äutomatedlydetermine the number of subgroups. The heterogeneity analysis of non-small-cell lung cancer based on single-cell gene expression data of the Wnt pathway and that of lung adenocarcinoma based on histopathological imaging data not only demonstrate the practical applicability of the proposed approach but also lead to interesting new findings.

Session 23INT21: Recent Developments for Dependent Data with Complex Structure

A Variational Bayesian Approach to Identifying Whole-Brain Directed Networks with Fmri Data

Yaotian Wang[1], Guofen Yan[2], Xiaofeng Wang[3], Shuoran Li[1], Lingyi Peng[1], Dana Tudorascu[1] and ♦Tingting Zhang[1]
[1]University of Pittsburgh
[2]University of Virginia
[3]Cleveland Clinic

The brain is a high-dimensional directed network systemas it consists of many regions as network nodes that exert influence on each other. The directed influence exerted by one region on another is referred to as directed connectivity. We aim to reveal whole-brain directed networks based on resting-state functional magnetic resonance imaging (fMRI) data of many subjects. However, it is both statistically and computationally challenging to produce scientifically meaningful estimates of whole-brain directed networks. To address the statistical modeling challenge, we assume modular brain networks, which reflect functional specialization and functional integration of the brain. We address the computational challenge by developing a variational Bayesian method to estimate the new model. We apply our method to resting-state fMRI data of many subjects and identify modules and directed connections in whole-brain directed networks. The identified modules are accordant with functional brain systems specialized for different functions.We also detect directed connections between functionally specialized modules, which is not attainable by existing network meth-

ods based on functional connectivity. In summary, this paper presents a new computationally efficient and flexible method for directed network studies of the brain as well as new scientific findings regarding the functional organization of the human brain.

Bayesian Image Mediation Analysis

Yuliang Xu and ♦Jian Kang
University of Michigan

Mediation analysis aims to separate the indirect effect through mediators from the direct effect of the exposure on the outcome. It is challenging to perform mediation analysis with neuroimaging data which involves high dimensionality, complex spatial correlations, sparse activation patterns and relatively low signal-to-noise ratio. To address these issues, we develop a new spatially varying coefficient structural equation model for Bayesian image mediation analysis (BIMA). Under the potential outcome framework, we formally define the spatially varying mediation effects of the exposure on the outcome that are mediated through imaging mediators. For prior specifications of spatially varying coefficients in BIMA, we adopt the soft-thresholded Gaussian process (STGP) which ensures a large prior support for sparse and piece-wise smooth functions. We establish and posterior consistency for spatially varying mediation effects along with selection consistency on important regions that contribute to the mediation effects. We develop an efficient posterior computation algorithm for BIMA which is scalable to analysis of large-scale imaging data. Through extensive simulations, we show that BIMA can improve the estimation accuracy and computational efficiency for high-dimensional mediation analysis over the existing methods. We apply BIMA to analyze the behavioral and fMRI data in the Adolescent Brain Cognitive Development (ABCD) study with a focus on inferring the mediation effects of the parental education level on the children's general cognitive ability that are mediated through the working memory brain activities.

Precision Education: a Bayesian Nonparametric Approach for Handling Item and Examinee Heterogeneity in Assessment Data

Tianyu Pan[1], Weining Shen[1], Clintin Stober[2] and ♦Guanyu Hu[2]
[1]University of California Irvine
[2]University of Missouri Columbia

We propose a novel nonparametric Bayesian IRT model in this paper by introducing the clustering effect at question level and further assume heterogeneity at examinee level under each question cluster, characterized by the mixture of Binomial distributions. The main contribution of this work is threefold: (1) We demonstrate that the model is identifiable. (2) The clustering effect can be captured asymptotically and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a $\sqrt{n}$ rate (up to a $\log$ term). (3) We present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. We evaluate our model via a series of simulations as well as apply it to an English assessment data. This data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

High-Dimensional Response Growth Curve Modeling for Longitudinal Neuroimaging Analysis

♦Lu Wang[1], Xiang Lyu[2], Zhengwu Zhang[3] and Lexin Li[2]
[1]Central South University
[2]University of California at Berkeley
[3]University of North Carolina at Chapel Hill

There is increasing interest in modeling high-dimensional longitudinal outcomes in applications such as developmental neuroimaging research. Growth curve model offers a useful tool to capture both the mean growth pattern across individuals, as well as the dynamic changes of outcomes over time within each individual. However, when the number of outcomes is large, it becomes challenging and often infeasible to tackle the large covariance matrix of the random effects involved in the model. In this article, we propose a high-dimensional response growth curve model, with three novel components: a low-rank factor model structure that substantially reduces the number of parameters in the large covariance matrix, a re-parameterization formulation coupled with a sparsity penalty that selects important fixed and random effect terms, and a computational trick that turns the inversion of a large matrix into the inversion of a stack of small matrices and thus considerably speeds up the computation. We develop an efficient expectation-maximization type estimation algorithm, and demonstrate the competitive performance of the proposed method through both simulations and a longitudinal study of brain structural connectivity in association with human immunodeficiency virus.

Session 23INT83: Statistical Learnning on Complex Data

Deep Kronecker Network

♦Long Feng[1] and Guang Yang[2]
[1]University of Hong Kong
[2]City University of HOng Kong

We propose Deep Kronecker Network (DKN), a novel framework designed for analyzing medical imaging data, such as MRI, fMRI, CT, etc. Medical imaging data is different from general images in at least two aspects: i) sample size is usually much more limited, ii) model interpretation is more of a concern compared to outcome prediction. Due to its unique nature, general methods, such as convolutional neural network (CNN), are difficult to be directly applied. As such, we propose DKN, that is able to adapt to low sample size limitation and provide desired model interpretation. DKN is general in the sense that it not only works for both matrix and (high-order) tensor represented image data, but also could be applied to both discrete and continuous outcomes. DKN is built on a Kronecker product structure and implicitly imposes a piecewise smooth property on coefficients. Moreover, the Kronecker structure can be written into a convolutional form, so DKN also resembles a CNN, particularly, a fully convolutional network (FCN). Interestingly, DKN is also highly connected to the tensor regression framework proposed by Zhou et al. (2013), where a CANDE-COMP/PARAFAC (CP) low-rank structure is imposed on tensor coefficients. We conduct both classification and regression analyses using real MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to demonstrate the effectiveness of DKN.

Fpls-Dc: Functional Partial Least Squares Through Distance Covariance for Imaging Genomics

♦Wenliang Pan[1], Chuang Li[2], Yue Shan[3], Tengfei Li[3], Yun Li[3] and Hongtu Zhu[3]

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[2]Sun Yat-sen university
[3]University of North Carolina at Chapel Hill

We proposed a functional partial least squares through distance correlation (FPLS-DC) framework with two components to efficiently carry out whole-genome wide analyses of functional phenotypes. The first one is to use the first FPLS-derived base function to reduce the dimentionality of image, while screening genetic markers. The second one is to maximize the distance correlation between genetic markers and projected imaging data, which is a linear combination of the first few FPLS-basis functions, through sequential quadratic programming. We use a gamma approximation to efficiently approximate the null distribution of test statistics. Compared with the existing estimation methods, FPLS- DC enjoys both computational efficiency and statistical efficiency for handling large-scale imaging genetics. In real application, the proposed approach successfully detected several novel Alzheimer's disease-related genetic variants and regions of interest, which indicate that our method may be a valuable statistical toolbox for imaging genetic study.

Ball Impurity: Measuring Heterogeneity in General Metric Spaces

Ting Li

The Hong Kong Polytechnic Universtiy

Various domains, such as neuroimaging and network data analysis, have data in complex forms which do not process a Hilbert structure. We propose ball impurity, a general measure of heterogeneity among complex non-Euclidean objects. Our approach measures the difference between distributions in general metric spaces by generalizing the impurity degree in Hilbert spaces. The measure has properties analogous to triangular inequalities, is straightforward to compute and can be used for variable screening and tree models.

Session 23INT68: Statistical Design and Analysis of Reliability and Survival Data

Kaplan-Meier Type Precedence Test Based on Ranked Set Progressively Type-Ii Censored Data

Chang Cui, ♦Tao Li, Jiaqi Men and Yijia Zheng

Shanghai University of Finance and Economics

In this article, we propose a Kaplan-Meier type precedence test statistics for testing the equality of two distributions, which is based on both conventional and progressively type-II censored ranked set samples. The exact null distribution of the proposed test statistic is derived, critical values are tabulated for different set size, number of cycle and progressively censoring scheme. The exact power functions under the Lehmann alternative for both conventional and progressively type-II censored ranked set samples are derived. The power values of the proposed test are exactly computed under the Lehmann alternative and also through Monte Carlo simulations under a location-shift alternative. To evaluate the proposed test statistic, the power values of the proposed test and some of competitors are compared under the conventional type-II censored ranked set samples.

Minimax Designs for Accelerated Life Tests

♦I-Chen Lee[1], Ray-Bing Chen[1] and Weng Kee Wong[2]

[1]National Cheng Kung University

[2]University of California, Los Angeles

Due to time constraint and experimental cost, how to plan an efficient accelerated life test (ALT) to obtain more accurate lifetime information of products is an important research issue. Many strategies were proposed to design a locally optimal planning of an ALT under the pre-specified planning values of parameters. However, the optimal design for an ALT also depends on model parameters are usually unknown before the experiment. To deal with the problems, this study adopts a minimax criterion to obtain a more robust design for conducting an ALT. Particularly, the minimax design is determined once we specify the range of sample failure probability under a specific failure model. To find the minimax design efficiently, this study adopts the particle swarm optimization (PSO) technique. Finally, compared to the locally optimal design via simulation study, the minimax design is more robust and more practical.

Reliability Estimation for One-Shot Devices with Correlated Components

Man Ho Ling

The Education University of Hong Kong

A device that performs its intended function only once is referred to as a one-shot device. Actual lifetimes of such kind of devices under life-tests cannot be observed. In addition, one-shot devices often consist of multiple components that could cause the failure of the device. The lifetime of each component under test is either left- or right-censored. The components are also coupled together in the manufacturing process or assembly, resulting in the failure modes possessing latent heterogeneity and dependence. Frailty models facilitate an easily understandable interpretation for the dependence between components. However, finding the maximum likelihood estimates of frailty models based on completely censored data is challenging. An efficient expectation–maximization algorithm is presented to find the maximum likelihood estimates of model parameters, on the basis of one-shot device testing data with multiple failure modes under a constant-stress accelerated life-test, with the dependent components having exponential lifetime distributions under gamma frailty. The maximum likelihood estimate and confidence intervals for the mean lifetime of k-out-of-M structured one-shot device under normal operating conditions are also discussed. The performance of the proposed inferential methods is finally evaluated through Monte Carlo simulations.

Session 23INT72: Recent Advancements in Statistical Methods for Complex Lifetime Data

Evaluation of the Natural History of Disease by Combining Incident and Prevalent Cohorts: Application to the Nun Study

♦Daewoo Pak[1], Jing Ning[2], Richard Kryscio[3] and Yu Shen[2]

[1]Yonsei University
[2]The University of Texas MD Anderson Cancer Center
[3]University of Kentucky

The Nun study is a well-known longitudinal epidemiology study of aging and dementia that recruited elderly nuns who were not yet diagnosed with dementia (i.e., incident cohort) and who had dementia prior to entry (i.e., prevalent cohort). In such a natural history of disease study, multistate modeling of the combined data from both incident and prevalent cohorts is desirable to improve the efficiency of inference. While important,

the multistate modeling approaches for the combined data have been scarcely used in practice because prevalent samples do not provide the exact date of disease onset and do not represent the target population due to left-truncation. In this paper, we demonstrate how to adequately combine both incident and prevalent cohorts to examine risk factors for every possible transition in studying the natural history of dementia. We adapt a four-state nonhomogeneous Markov model to characterize all transitions between different clinical stages, including plausible reversible transitions. The estimating procedure using the combined data leads to efficiency gains for every transition compared to those from the incident cohort data only.

Interval-Censored Linear Rank Regression

♦Sangbum Choi[1], Taehwa Choi[2] and Wenbin Lu[3]
[1]Department of Statistics, Korea University
[2]Department of Biostatistics and Bioinformatics, Duke University
[3]Department of Statistics, North Carolina State University

This paper introduces a novel rank-based approach to make inferences about the regression parameter in the semiparametric accelerated failure time model with interval-censored and current status data. This type of data can arise when failure times are not exactly known, but are known only to have occurred between some intermittent monitoring times. A broad class of monotone log-rank estimating equations, which does not involve computing the nonparametric maximum likelihood estimate of the baseline distribution function at the residuals, is constructed for parameter estimation. The corresponding estimators can be immediately obtained via linear programming and are shown to be consistent and asymptotically normal. A numerically efficient resampling procedure is considered for confidence procedure. Furthermore, an one-step adaptive procedure based on estimating the optimal weights of the log-rank equation is explored to improve statistical efficiency. Extensive numerical studies are conducted to evaluate the finite-sample performance of the new estimators. Our approach is illustrated with two data examples from HIV/AIDS cohort studies.

Quantile Association Regression on Bivariate Survival Data

Ling-Wan Chen[1], ♦Yu Cheng[2], Ying Ding[2] and Ruosha Li[3]
[1]FDA
[2]University of Pittsburgh
[3]UT Health

The association between two event times is of scientific importance in various fields. Due to population heterogeneity, it is desirable to examine the degree to which local association depends on different characteristics of the population. Here we adopt a novel quantile-based local association measure and propose a conditional quantile association regression model to allow covariate effects on local association of two survival times. Estimating equations for the quantile association coefficients are constructed based on the relationship between this quantile association measure and the conditional copula. Asymptotic properties for the resulting estimators are rigorously derived, and induced smoothing is used to obtain the covariance matrix. Through simulations we demonstrate the good practical performance of the proposed inference procedures. An application to age-related macular degeneration (AMD) data reals interesting varying effects of the baseline AMD severity score on the local association between two AMD progression times.

Session 23INT77: Recent Advances in Nonparametric Methods in Time Series and Econometrics

Mean Stationarity Test in Time Series: a Signal Variance-Based Approach

♦Hon Kiu To and Kin Wai Chan
The Chinese University of Hong Kong

Inference of mean structure is an important problem in time series analysis. Various tests have been developed to test for different mean structures, including but not limited to, the presence of structural break(s), and parametric mean structures. Many of them are designed under specific mean structures, and may potentially lose power upon violation of such structures. We propose a new mean stationarity test built around the signal variance. The proposed test can detect the non-constancy of the mean function under serial dependence. It is shown to have promising power in detecting hardly noticeable periodic structures. The proposal is further generalized to test for smooth mean structures and relative signal variability. A real-data application on global land surface temperature data is presented.

Optimally Jittered Jump Test for High-Frequency Data

Cheuk Hin Cheng, Hon Kiu To, Kai Pan Chu and ♦Ting Tin Ma
Department of Statistics, CUHK

Bipower variation is an important quantity in finance economics, It can be used to estimate integrated volatility of the process and detect existence of potential jumps. We propose the concept of auto-bipower variation, which is a generalization of bipower variation. It serves as building block to construct more precise estimators of the integrated stochastic volatility and a series of feasible tests on the detection of potential jumps. Jittering the those tests with higher-order multi-power variation further yields an optimally jittered test, which is the most powerful in the proposed tests and it does not rely on the tuning of any hyper-parameter.

A Non-Parametric Approach for Causal Inference in Time Series

♦Kai Pan Chu and Kin Wai Chan
Department of Statistics, The Chinese University of Hong Kong

This paper addresses the problem of estimating the treatment effect of an action on a time series. The causality is understood in a modified Rubin's framework of potential outcomes. The variance estimator is non-trivial in the sense that the standard estimator for asymptotic variance in time series fails to be consistent in the situation of causal inference. Some existing hypothesis testing procedure for drawing causal inference has been shown to be too conservative in many situations, especially when the mean of the series is far away from zero. In the presentation, we give a precise definition for various types of treatment effects. A fully non-parametric setting is considered. For the time-average effect, we propose a Horvitz-Thompson type estimator and derive a consistent estimator for its variance. The powerfulness of the test based on our proposed estimators is guaranteed without making any parametric or time-independent assumption. We also propose a Nadaraya-Watson type estimator and provide a bootstrap-type simultaneous confidence band procedure for the time-varying effect. Monte Carlo experiments are included to demonstrate the finite-sample performance of our proposal.

General Framework for Self-Normalized Multiple-Change-Point Tests

♦*Cheuk Hin Cheng and Kin Wai Chan*

We propose a general framework to construct self-normalized multiple-change-point tests with time series data. The only building block is a user-specified single-change-detecting statistic, which covers a large class of popular methods, including the cumulative sum process, outlier-robust rank statistics, and order statistics. The proposed test statistic does not require robust and consistent estimation of nuisance parameters, selection of bandwidth parameters, nor pre-specification of the number of change points. The finite-sample performance shows that the proposed test is size-accurate, robust against misspecification of the alternative hypothesis, and more powerful than existing methods. Case studies of the Shanghai-Hong Kong Stock Connect turnover are provided.

Session 23INT79: Semiparametric Inference for Complex Data

Multiple Descent in Random Feature Regression

Xuran Meng[1], Jianfeng Yao[2] and ♦Yuan Cao[1]
[1]The University of Hong Kong
[2]The Chinese University of Hong Kong (Shenzhen)

Recent research has revealed a double descent phenomenon in over-parameterized regression models, where the excess risk initially decreases, then increases, and then decreases again as the model complexity increases. Although this phenomenon has been explored in various settings, it is not fully understood in theory. An open question is whether similar phenomena occur in more complex models comprising multiple components. In this talk, I will present an investigation of this problem through the lens of random feature regression. Specifically, we study the excess risk of a double random feature model (DRFM) consisting of two types of random features in ridge regression. We derive the precise limit of the excess risk under the high-dimensional framework, where the training sample size, the dimension of data, and the dimension of random features tend to infinity proportionally. Our analysis reveals that the risk curves of DRFMs can exhibit triple descent, providing further insights into this phenomenon. We further extend our analysis to multiple random feature models (MRFMs) and demonstrate that MRFMs with K types of random features may exhibit (K+1)-fold descent. Our study suggests that risk curves with a specific number of descent generally exist in random feature regression, and sheds light on the multiple descent phenomenon in more complex models such as neural network models and semi-parametric regression models.

Phase-Type Sieve

Zhisheng Ye
National University of Singapore

In many nonparametric and semiparametric models, the infinite-dimensional parameter of direct interest is the probability density, but its nonparametric estimation is usually difficult in the presence of incomplete data. In this talk, we design a sieve class of phase-type densities whose limit is dense in the class of nonnegative densities, and we establish its approximation error rate for a given density. The proposed phase-type sieve is then used in M-estimation for a semiparametric model, in which the

nonparametric component is a density. We demonstrate our method using several examples.

Temporal Heterogeneity Learning for Functional Panel Quantile Regressions

♦*Jiaqi Men[1], Jinhong You[1] and Hua Liu[2]*
[1]Shanghai University of Finance and Economics
[2]Xi'an Jiaotong University

A partial functional quantile regression model for panel data with time-varying parameters is proposed to investigate the temporal heterogeneity. The function- valued parameter in the proposed model is assumed to change over unknown time regimes. A three-step estimation technique to estimate the completely time-varying parameter vector, regime-varying function-valued parameter, as well as the unknown time regimes, is developed. In order to cope with the computational complexity posed by the large sample and high-dimensional characteristics of the panel and functional data, we apply a convolution-type smoothing approach conquer proposed by He et al. (2021) to smooth the objective function. Asymptotical normality of the resultant es- timators and consistency of the identification procedure used to detect unknown time regimes are established. The simulation study and two real applications of the "id- iosyncratic volatility puzzle" and air pollution illustrate the competitive performance of the proposed model and the corresponding statistical inference methods.

Session 23INT87: Network Structure and Structural Change-Point Estimation

Sparse Change Detection in High-Dimensional Linear Regression

Fengnan Gao[1] and ♦Tengyao Wang[2]
[1]Fudan University
[2]LSE

We introduce a new methodology 'charcoal' for estimating the location of sparse changes in high-dimensional linear regression coefficients, without assuming that those coefficients are individually sparse. The procedure works by constructing different sketches (projections) of the design matrix at each time point, where consecutive projection matrices differ in sign in exactly one column. The sequence of sketched design matrices is then compared against a single sketched response vector to form a sequence of test statistics whose behaviour shows a surprising link to the well-known CUSUM statistics of univariate change-point analysis. Strong theoretical guarantees are derived for the estimation accuracy of the procedure, which is computationally attractive, and simulations confirm that our methods perform well in a broad class of settings.

Changepoint Detection in Preferential Attachment Networks

Daniel Cirkovic[1], ♦Tiandong Wang[2] and Xianyang Zhang[1]
[1]Texas A&M University
[2]Fudan University

Generative, temporal network models play an important role in analyzing the dependence structure and evolution patterns of complex networks. Due to the complicated nature of real network data, it is often naive to assume that the underlying data-generative mechanism itself is invariant with time. Such observation leads to the study of changepoints or sudden shifts

in the distributional structure of the evolving network. We propose both likelihood-based and extreme-value-based methods to detect changepoints in undirected, affine preferential attachment networks, and establish a hypothesis testing framework to detect a single changepoint, together with a consistent estimator for the changepoint.

Community Detection in Sparse Latent Space Models
♦Fengnan Gao[1], Hongsong Yuan and Zongming Ma
[1]Fudan University

We show that a simple community detection algorithm originated from stochastic blockmodel literature achieves consistency, and even optimality, for a broad and flexible class of sparse latent space models. The class of models includes latent eigenmodels (Hoff, 2008). The community detection algorithm is based on spectral clustering followed by local refinement via normalized edge counting. It is easy to implement and attains high accuracy with a low computational budget. The proof of its optimality depends on a neat equivalence between likelihood ratio test and edge counting in a simple vs. simple hypothesis testing problem that underpins the refinement step, which could be of independent interest.

Root and Community Inference on Markovian Models of Networks
Min Xu
Rutgers University

Many existing statistical models for networks overlook the fact that most real-world networks are formed through a growth process. To address this, we introduce the PAPER (Preferential Attachment Plus Erdos-Renyi model for random networks, where we let a random network G be the union of a preferential attachment (PA) tree T and additional Erdos-Renyi (ER) random edges. The PA tree component captures the underlying growth/recruitment process of a network where vertices and edges are added sequentially, while the ER component can be regarded as random noise. Given only a single snapshot of the final network G, we study the problem of constructing confidence sets for the early history, in particular the root node, of the unobserved growth process; the root node can be patient zero in a disease infection network or the source of fake news in a social media network. We propose an inference algorithm based on Gibbs sampling that scales to networks with millions of nodes and provide theoretical analysis showing that the expected size of the confidence set is small so long as the noise level of the ER edges is not too large. We also propose variations of the model in which multiple growth processes occur simultaneously, reflecting the growth of multiple communities, and we use these models to provide a new approach to community detection.

Session 23INTSP2: Special Memorial Session to Celebrate Life of Professor Tze Leung Lai

Tba
Gang Li
UCLA
TBA

Advancing Sequential Importance Sampling: a Tribute to Professor Tze Leung Lai
Yuguo Chen
University of Illinois Urbana-Champaign

In this talk, we highlight the contributions of Professor Tze Leung Lai to sequential importance sampling (SIS) methods and their applications. SIS is a versatile and powerful Monte Carlo method for tackling complex statistical inference problems. Professor Lai's work conducted rigorous theoretical analyses of SIS and provided deep insights into the development of more effective SIS methods. His work has significantly enriched the theory and practice of SIS, and will continue to influence the field of sequential Monte Carlo methods and provide a foundation for future research in simulation-based inference.

Innovations in Clinical Trial Methodology for Precision Medicine: a Tribute to Professor Tze Leung Lai
♦Ying Lu and Lu Tian
Stanford University

This talk aims to pay tribute to Professor Tze Leung Lai's significant contributions, honoring his legacy as a visionary researcher and educator. We will reflect on his innovations in clinical trial methodology, his unwavering dedication to precision medicine, and his relentless pursuit of advancing statistical science in the realm of drug development. We will highlight some of his notable innovations in clinical trial methodology that have significantly enhanced the efficiency and accuracy of these trials. Furthermore, we will discuss Professor Lai's remarkable efforts to bridge the gap between statistical research and practical implementation through the establishment and leadership of the Stanford Center for Innovative Study Design.

Professor Lai's Contributions to Sequential Experimentation
Zhiliang Ying
Columbia University

Professor Lai was a leading authority in and made seminal contributions to the sequential methodology. This talk summarizes his works on nonlinear renewal theory, stochastic approximation, adaptive control, multi-armed bandit problem, and group sequential methodology among others. Their connections to the modern-age statistics are also discussed.

Professor Lai's Contributions and Influence on Statistical Research in Taiwan
♦Ching-Kang Ing[1], I-Ping Tu[2] and Chao A. Hsiung[3]
[1]National Tsing Hua University
[2]Academia Sinica, Taiwan
[3]National Health Research Institutes, Taiwan

Professor Lai was an extraordinary individual who profoundly connected with Taiwan's statistical society. We aim to honor Professor Lai's remarkable influence on Taiwan's statistical research during this presentation, highlighting his fruitful partnerships with local scholars. We sincerely appreciate Professor Lai's contributions and impact in the field, which will be forever cherished. His steadfast commitment, vast knowledge, and mentorship have left an indelible impression, and we genuinely hope his memory will continue inspiring future generations of statisticians.

Session 23INT109: Complex Data Analysis

Covariate Balancing with Measurement Error
Ying Yan
Sun Yat-sen University

In the past decade, there is an emerging literature on developing covariate balancing methods among statisticians and applied re-

searchers, where covariate balance is directly incorporated in the estimation procedure. It has been well documented that covariate balancing is superior to propensity score weighting in many circumstances. The validity of covariate balancing requires implicitly that all covariates are accurately measured. Measurement error is ubiquitous in real studies, but there is a lack of understanding of the role of measurement error in covariate balancing. In this talk, we systematically study the impact of measurement error on covariate balancing methods. In theory, we show that naive covariate balancing that ignores measurement error results in biased causal effect estimation and poor balancing performance. We then propose a class of measurement error correction strategies that successfully remove measurement error bias and achieve valid causal conclusions. Finally, we apply the proposed methods in simulation studies and the analysis of a lifetime data set.

Estimation and Model Selection for Nonparametric Function-on-Function Regression

◆*Zhanfeng Wang*[1], *Hao Dong*[2], *Ping Ma*[3] *and Yuedong Wang*[2]
[1]University of Science and Technology of China
[2]University of California, Santa Barbara
[3]University of Georgia
Regression models with a functional response and functional covariate have received significant attention recently. While various nonparametric and semiparametric models have been developed, there is an urgent need for model selection and diagnostic methods. In this article, we develop a unified framework for estimation and model selection in nonparametric function-on-function regression. We propose a general nonparametric functional regression model with the model space constructed through smoothing spline analysis of variance (SS ANOVA). The proposed model reduces to some of the existing models when selected components in the SS ANOVA decomposition are eliminated. We propose new estimation procedures under either L_1 or L_2 penalty and show that the combination of the SS ANOVA decomposition and L_1 penalty provides powerful tools for model selection and diagnostics. We establish consistency and convergence rates for estimates of the regression function and each component in its decomposition under both the L_1 and L_2 penalties. Simulation studies and real examples show that the proposed methods perform well. Technical details and additional simulation results are available in online supplementary materials.

Deep Image-on-Scalar Regression Model with Hidden Confounders

◆*Xiaohe Chen, Lintao Tang, Rongjie Liu and Chao Huang*
Florida State University
Integrative regression analysis with imaging responses from multiple sources usually suffers from heterogeneity caused by differences in study design, protocol, environment, population, or other hidden confounders. To address this challenge, this paper proposes a deep learning-based image-on-scalar regression model, which can simultaneously detect hidden confounders through surrogate variable analysis and primary effects from variables of interest using deep neural networks. Compared to existing solutions, the proposed method successfully handles imaging heterogeneities and captures complex association patterns. We establish both estimation and inference procedures for unknown varying coefficients and hidden confounders in our model. The asymptotic properties of the estimation procedure

are systematically investigated. The finite-sample performance of our proposed method is assessed by using both Monte Carlo simulations and a real data example on knee femoral MRI images from the osteoarthritis initiative (OAI) database.

Session 23INT91: Recent Developments in Biostatistics with their Applications

Recent Developments in Biostatistics with their Applications

Chen Lyu
Department of Population Health, NYU Grossman School of Medicine
TBA

Accurate Estimation of Breakpoints in Piecewise Linear Mixed-Effects Models with Application to Longitudinal Ophthalmic Studies

Jiyuan Hu
NYU Grossman School of Medicine
Purpose: Broken stick analysis is a widely used approach for detecting unknown breakpoints where association between measurements is non-linear. Existing longitudinal ophthalmic studies aggregate measurements obtained from all visits without considering the repeated measurements from a given eye so that segmented linear models can be applied to such "compressed" cross-sectional data. The purpose of this study is to propose an advanced robust segmented mixed-effects model (RSMM) which accommodates longitudinal measurements from both eyes and is robust to outliers. Methods: The model setup of RSMM and computing algorithm for point and confidence interval estimates of the breakpoint was introduced. The performance of RSMM and other competing methods was assessed via comprehensive simulation studies and application to a longitudinal ophthalmic study with 216 eyes (145 subjects) followed for an average of 3.7 1.3 years to examine the longitudinal association between structural and functional measurements. Results: In simulation studies, RSMM showed the smallest bias and mean squared error (MSE) for estimating the breakpoint, with empirical coverage probability of corresponding CI estimate closest to the nominal level for scenarios with and without outlier data points. In the application to the longitudinal ophthalmic study, RSMM detected two breakpoints between visual field mean deviation (MD) and retinal nerve fiber layer thickness (RNFL) and one breakpoint between MD and cup to disc ratio (CDR), while the cross-sectional analysis approach only detected one and none, respectively. Conclusions: RSMM improves the estimation accuracy of breakpoints for longitudinal ophthalmic studies. The conventional cross-sectional analysis approach is not recommended for future studies.

Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-Omics Data, and Predicting Disease Risk

◆*Chan Wang, Leopoldo Segal, Jiyuan Hu, Boyan Zhou, Richard Hayes, Jiyoung Ahn and Huilin Li*
New York University Grossman School of Medicine
Background With the rapid accumulation of microbiome-wide association studies, a great amount of microbiome data are available to study the microbiome's role in human disease and advance the microbiome's potential use for disease prediction.

However, the unique features of microbiome data hinder its utility for disease prediction. Methods Motivated from the polygenic risk score framework, we propose a microbial risk score (MRS) framework to aggregate the complicated microbial profile into a summarized risk score that can be used to measure and predict disease susceptibility. Specifically, the MRS algorithm involves two steps: (1) identifying a sub-community consisting of the signature microbial taxa associated with disease and (2) integrating the identified microbial taxa into a continuous score. The first step is carried out using the existing sophisticated microbial association tests and pruning and thresholding method in the discovery samples. The second step constructs a community-based MRS by calculating alpha diversity on the identified sub-community in the validation samples. Moreover, we propose a multi-omics data integration method by jointly modeling the proposed MRS and other risk scores constructed from other omics data in disease prediction. Results Through three comprehensive real-data analyses using the NYU Langone Health COVID-19 cohort, the gut microbiome health index (GMHI) multi-study cohort, and a large type 1 diabetes cohort separately, we exhibit and evaluate the utility of the proposed MRS framework for disease prediction and multi-omics data integration. In addition, the disease-specific MRSs for colorectal adenoma, colorectal cancer, Crohn's disease, and rheumatoid arthritis based on the relative abundances of 5, 6, 12, and 6 microbial taxa, respectively, are created and validated using the GMHI multi-study cohort. Especially, Crohn's disease MRS achieves AUCs of 0.88 (0.85–0.91) and 0.86 (0.78–0.95) in the discovery and validation cohorts, respectively. Conclusions The proposed MRS framework sheds light on the utility of the microbiome data for disease prediction and multi-omics integration and provides a great potential in understanding the microbiome's role in disease diagnosis and prognosis.

Session 23INT92: Incorporating External Data in Superiority and Non-Inferiority Clinical Trials: Bayesian Nonparametric vs Parametric Models

A Bayesian Nonparametric Model for External Data with Application to Clinical Trials

♦Dehua Bi and Yuan Ji
The University of Chicago

We consider a new Bayesian nonparametric model that borrows information across external data sets with an aim enhance the efficiency of an ongoing clinical trial. The model, named Shared Atoms Model (SAM), induces a dependence clustering structure across grouped data sets. Each atom, i.e., cluster, represents a homogeneous subpopulation of patients that are believed to respond to a treatment or control arm similarly. SAM automatically infers commons atoms as well as unique atoms for each data set so that information from common atoms can be shared. In this fashion, the efficiency of a clinical trial may be improved if the patient samples in the trial share common clusters with external data. Applications of SAM include augmenting or synthesizing a control arm using external data. Examples will be provided.

A Bayesian Parametric Model to Incorporate Real-World Evidence in Pragmatic Trials

♦Jun Yin, Peter Noseworthy and Xiaoxi Yao

Mayo Clinic
Randomized clinical trials (RCTs) are considered the standard approach for assessing drug efficacy; however, patient recruitment in RCTs can be challenging for rare diseases. Innovative approaches, such as information borrowing, which leverages external control data can enhance a clinical trial's efficiency. Herein, we propose an innovative external control design, which augments the normal control arm with the real-world control. The method allows adaptive borrowing. Performance is evaluated using simulation studies motivated by a real-life pragmatic trial.

Session 23INT22: Advances in Statistical Genetics and Genomics

Funcode: Scoring Cross-Species Functional Conservation of Dna Elements using Encode Data

Weixiang Fang[1], Chaoran Chen[2], Boyang Zhang[1], Yi Wang[1], Ruzhang Zhao[1], Weiqiang Zhou[1] and ♦Hongkai Ji[1]
[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[2]Department of Biomedical Engineering, Johns Hopkins University

Evolutionary conservation is an important tool for identifying functional DNA elements in genomes and provides a foundation for studying human diseases using animal models. Conservation in DNA sequences does not necessarily imply conservation in dynamic functional activities. Quantifying functional conservation, however, has been constrained by limited availability of functional genomic data from matching samples across species. Here we present FUNCODE, a solution to scoring functional conservation of DNA elements by integrating data across species without requiring manually matched samples. By using computational sample matching, FUNCODE more accurately scores functional conservation and offers scalability to new samples and ability to score different data modalities. As part of the Encyclopedia of DNA Elements (ENCODE) efforts, we systematically scored human-mouse conservation of DNA regulatory elements based on chromatin accessibility and histone modifications. We further demonstrate utility of FUNCODE in finding new cis-regulatory elements, identifying discoveries translatable across species, and cross-species single-cell genomic data integration.

Probabilistic Cell/Domain-Type Assignment of Spatial Transcriptomics Data with Spatialanno

Xingjie Shi
East China Normal University

In the analysis of both single-cell RNA sequencing (scRNA-seq) and spatially resolved transcriptomics (SRT) data, classifying cells/spots into cell/domain types is an essential analytic step for many secondary analyses. Most of the existing annotation methods have been developed for scRNA-seq datasets without any consideration of spatial information. Here, we present SpatialAnno, an efficient and accurate annotation method for spatial transcriptomics datasets, with the capability to effectively leverage a large number of non-marker genes as well as "qualitative" information about marker genes without using a reference dataset. Uniquely, SpatialAnno estimates low-dimensional embeddings for a large number of non-marker genes via a factor model while promoting spatial smoothness among

neighboring spots via a Potts model. Using both simulated and four real spatial transcriptomics datasets from the 10x Visium, ST, Slide-seqV1/2, and seqFISH platforms, we showcase the method's improved spatial annotation accuracy, including its robustness to the inclusion of marker genes for irrelevant cell/domain types and to various degrees of marker gene misspecification. SpatialAnno is computationally scalable and applicable to SRT datasets from different platforms. Furthermore, the estimated embeddings for cellular biological effects facilitate many downstream analyses.

Evaluation of Epitranscriptome-Wide N6-Methyladenosine Di erential Analysis Methods

Daoyu Duan[1], Wen Tang[1], Runshu Wang[2], ♦Zhenxing Guo[3] and Hao Feng[1]
[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University
[2]Department of Biostatistics, University of Michigan
[3]School of Data Science, The Chinese University of Hong Kong - Shenzhen

RNA methylation has emerged recently as an active research domain to study post-transcriptional alteration in gene expression regulation. Various types of RNA methylation, especially N6-methyladenosine (m6A), are involved in human disease development. One of the fundamental questions in RNA methylation data analysis is to identify the Differentially Methylated Regions (DMRs), by contrasting cases and controls. Multiple statistical approaches have been recently developed for m6A DMR detection, but there is a lack of a comprehensive evaluation for these analytical methods. Here, we thoroughly assess all eight existing methods for DMR calling, using both synthetic and real data. Our simulation adopts a Gamma-Poisson model and logit linear framework, and accommodates various sample sizes and DMR proportions for benchmarking. For all methods, low sensitivities are observed among regions with low input levels, but they can be drastically boosted by an increase in sample size. TRESS and exomePeak2 perform the best using metrics of detection precision, FDR, type I error control, and runtime, though hampered by low sensitivity. Analyses on three real datasets suggest differential preference on identifed DMR length and uniquely discovered regions, between these methods.

Rna Velocity Estimation with Stochastic Differential Equations

♦*Xu Liao[1], Lican Kang[1], Xiaoran Chai[1], Yuling Jiao[2] and Jin Liu[3]*
[1]National University of Singapore
[2]Wuhan University
[3]The Chinese University of Hong Kong, Shenzhen

Recently, RNA velocity has driven a shift of paradigm in single-cell RNA sequencing (scRNA-seq) studies, enabling us to predict the developmental trajectory of individual cells and gain deeper insights into the mechanisms governing cellular fate. However, existing methods based on ordinary differential equations (ODEs) may be limited to fully capture the stochastic nature of transcription dynamics. Here, we present SDEvelo, a novel deep learning approach to solve the RNA velocity problem by establishing and solving stochastic differential equations (SDEs). SDEvelo explicitly models the inherent noise and uncertainty in transcription dynamics and identifies the parameters of SDEs by matching the distribution between real data and data generated by the model. Using both simulated and real data, we show

the improved accuracy of inferred differentiation directions and estimated RNA velocity can facilitate many downstream analyses. We demonstrate SDEvelo is computational scalable and applicable to scRNA-seq data from different platforms.

Session 23INT9: Bayesian Spatial Analysis: Theory, Method, and Application

Bayesian Fixed-Domain Asymptotics for Covariance Parameters in Spatial Gaussian Process Regression Models

♦*Cheng Li[1], Saifei Sun[1] and Yichen Zhu[2]*
[1]National University of Singapore
[2]Duke University

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.

Bayesian Modeling of Spatially Resolved Transcriptomics Data

Qiwei Li
The University of Texas at Dallas

The location, timing, and abundance of gene expression within a tissue define the molecular mechanisms of cell functions. Recent technology breakthroughs in spatial molecular profiling, including imaging-based technologies and sequencing-based technologies, have enabled the comprehensive molecular characterization of single cells while preserving their spatial and morphological contexts. This new bioinformatics scenario calls for effective and robust computational methods to identify genes with spatial patterns. We represent two novel Bayesian hierarchical models to analyze spatial molecular profiling data, with several unique characteristics. The first model based on Gaussian process directly models the zero-inflated and over-dispersed counts. The second model based on Ising model uses the energy interaction parameter to characterize a denoised spatial pattern. The Bayesian inference framework allows us to borrow strength in parameter estimation in a de novo fashion. The two proposed models show competitive performances in accuracy and robustness over existing methods in both simulation studies and two real data applications.

Characterizing the Extremal Dependence in Spatial Analysis

of 2021 Pacific Northwest Heatwave

♦*Likun Zhang*[1], *Mark Risser*[2], *Michael Wehner*[2] and *Travis O'brien*[3]

[1] University of Missouri
[2] Lawrence Berkeley National Laboratory
[3] Indiana University

In late June, 2021, a devastating heatwave affected the US Pacific Northwest and western Canada, breaking numerous all-time temperature records by large margins and directly causing hundreds of fatalities. This unprecedented event was unforeseeable even after accounting for anthropogenic climate change, making it impossible to explain its abnormality or quantify the probability of a similar event in the future. Furthermore, the observed 2021 daily maximum temperature across much of the Pacific Northwest exceeded upper bound estimates obtained from single-station temperature records, meaning that the event could not have been predicted under standard univariate extreme value analysis assumptions. In this work, we utilize a flexible spatial extremes model that considers all stations across the Pacific Northwest domain and accounts for the fact that many stations simultaneously experience extreme temperatures. Our analysis incorporates the effects of anthropogenic forcing and natural climate variability in order to better characterize time-varying changes in the distribution of daily temperature extremes. We show that greenhouse gas forcing, drought conditions and large-scale atmospheric modes of variability all have significant impact on summertime maximum temperatures in this region. Nonetheless, while our model represents a significant improvement over corresponding single-station analysis, we are unable to fully anticipate the observed 2021 high temperatures even after properly accounting for extremal dependence, reiterating the uniqueness and unpredictability of the 2021 heatwave in the Pacific Northwest.

Uncertainty Quantification for Object Boundaries Extracted from Spatial Point Patterns

Jue Wang[1], *Vinay Kashyap*[2], ♦*Thomas Lee*[1], *David Van Dyk*[3] and *Andreas Zezas*[4]

[1] UC Davis, USA
[2] Center for Astrophysics | Harvard & Smithsonian, USA
[3] Imperial College London, UK
[4] deUniversity of Crete, Greece

A common challenge in astronomy lies in the reconstruction of astronomical objects based on photon counts detected by telescopes, as well as in determining whether these objects exhibit temporal variations. This research aims to address these issues by developing methods to unveil obscured objects within noisy spatial point patterns, quantifying uncertainties associated with the boundaries of these objects, and ultimately identifying changes that may have occurred over time. In this talk, we present our ongoing efforts toward solving this problem, highlighting our novel approaches and their potential impact on astronomy.

Session 23INT73: Challenges and Advances in Risk Assessment and Prediction

Conditional Concordance-Assisted Learning for Combining Biomarkers for Population Screening

♦*Wen Li*[1], *Ruosha Li*[2], *Qingxiang Yan*[3], *Ziding Feng*[4] and *Jing Ning*[5]

[1] The University of Texas McGovern Medical School at Houston, TX, USA
[2] The University of Texas School of Public Health, TX, USA
[3] F. Hoffmann-La Roche Ltd., ON, Canada
[4] Fred Hutchinson Cancer Research Center, WA, USA
[5] The University of Texas MD Anderson Cancer Center, TX, USA

Incorporating promising biomarkers into cancer screening practices for early-detection is increasingly appealing because of the unsatisfactory performance of current cancer screening strategies. The matched case-control design is commonly adopted in biomarker development studies to evaluate the discriminative power of biomarker candidates, with an intention to eliminate confounding effects. Data from matched case-control studies have been routinely analyzed by the conditional logistic regression, although the assumed logit link between biomarker combinations and disease risk may not always hold. We propose a conditional concordance-assisted learning method, which is distribution-free, for identifying an optimal combination of biomarkers to discriminate cases and controls. We are particularly interested in combinations with a clinically and practically meaningful specificity to prevent disease-free subjects from unnecessary and possibly intrusive diagnostic procedures, which is a top priority for cancer population screening. We establish asymptotic properties for the derived combination and confirm its favorable finite sample performance in simulations. We apply the proposed method to the prostate cancer data from the Carotene and Retinol Efficacy Trial (CARET).

Semiparametric Isotonic Regression Model and Estimation for Group Testing Data

Ao Yuan[1], ♦*Jin Piao*[2], *Jing Ning*[3] and *Jing Qin*[4]

[1] Georgetown University
[2] The University of Southern California
[3] The University of Texas MD Anderson Cancer Center
[4] National Institute of Allergy and Infectious Diseases

In the group testing procedure, several individual samples are grouped and the pooled samples, instead of each individual sample, are tested for outcome status (e.g., infectious disease status). Although this cost-effectiveness strategy in data collection is both labour and time-efficient, it poses statistical challenges to derive statistically and computationally efficient estimators under semiparametric models. We consider semiparametric isotonic regression models for the simultaneous estimation of the conditional probability curve and covariate effects, in which a parametric form for combining the covariate information is assumed and the monotonic link function is left unspecified. We develop an expectation–maximization algorithm to overcome the computational challenge and embed the pool-adjacent violators algorithm in the M-step to facilitate the computation. We establish the large sample behaviour of the proposed estimators and examine their finite sample performance in simulation studies. We apply the proposed method to data from the National Health and Nutrition Examination Survey for illustration.

Analysis of Survival Data with Cure Fraction and Variable Selection: a Pseudo-Observations Approach

Chien-Lin Su[1], ♦*Sy Han Chiou*[2], *Feng-Chang Lin*[3] and *Robert Platt*[1]

[1] McGill University
[2] University of Texas at Dallas
[3] University of North Carolina

The mixture cure model for analyzing survival data with cure fraction consists of an incidence component that indicates whether the subject is cured and a latency component that models the time to the event among non-cured patients. We propose a new estimating procedure that utilizes the novel pseudo-observations approach to bypass the computationally intensive EM algorithm required in standard approaches. The proposed pseudo-observations approach also allows researchers to assess covariate effects and perform variable selection in each incidence and latency component separately. The proposed method is extended to the bounded cumulative hazard model (promotion time cure model). Extensive simulation studies indicate superior performance of the proposed estimators over existing methods. An informal goodness-of-fit assessment is proposed based on pseudo-residuals to provide guidelines for selecting between the mixture cure model and the bounded cumulative hazard model. The proposed method is demonstrated through applications.

Session 23INT93: Showcase of the Power of Statistics in Observational Studies for Precision Health

Statistical Opportunities in Analyzing Real-World Interventional Mobile Health Data
Zhenke Wu
University of Michigan, Ann Arbor
Twin revolutions in wearable technologies and smartphone-delivered digital health interventions have significantly expanded the accessibility and uptake of personalized interventions in multiple domains of health sciences. For example, push notifications to promote healthy behaviors can be sent via mobile device that are adapted to continuously collected information on an individual's current context. These time-varying adaptive interventions are hypothesized to lead to meaningful short and long-term behavior change. This talk will formulate key scientific questions in statistical terms. However, standard assumptions such as non-interference and stationarity might be violated in real-world mobile health studies due to peer influence and long monitoring periods. I will present two methodological solutions, the first for estimating a new type of peer effects and the second for optimal policy learning under possibly non-stationary environments. I will use a multi-institution cohort of first year medical interns in the United States to illustrate the ideas. I will highlight that teams of engineers, clinical and data scientists can collaborate to build statistical models that extract scientific insights from noisy and longitudinal interventional mobile health data.

Constructing Time-Invariant Dynamic Surveillance Rules for Optimal Monitoring Schedules
Yingqi Zhao
Dynamic surveillance rules (DSRs) are sequential surveillance decision rules informing monitoring schedules in clinical practice, which can adapt over time according to a patient's evolving characteristics. In many clinical applications, it is desirable to identify and implement optimal time-invariant DSRs, where the parameters indexing the decision rules are shared across different decision points. We propose a new criterion for DSRs that accounts for benefit-cost tradeoff during the course of disease surveillance. We develop two methods to estimate the time-invariant DSRs optimizing the proposed criterion, and establish asymptotic properties for the estimated parameters of biomarkers indexing the DSRs. The first approach estimates the optimal

decision rules for each individual at every stage via regression modeling, and then estimates the time-invariant DSRs via a classification procedure with the estimated time-varying decision rules as the response. The second approach proceeds by optimizing a relaxation of the empirical objective, where a surrogate function is utilized to facilitate computation. Extensive simulation studies are conducted to demonstrate the superior performances of the proposed methods. The methods are further applied to the Canary Prostate Active Surveillance Study (PASS)

Reinforcement Learning for Estimating Optimal Dynamic Treatment Rules
Weijie Liang and ♦Jinzhu Jia
Peking University, China
Dynamic treatment rules (DTR) are very import for a patient to receive a treatment. A patient could gain enough benefits when a good dynamic treatment rule is applied. A dynamic treatment rule is a sequence of decision rules that depends on a sequence of the states of one patient. Recently, Reinforcement learning has gained great success in estimating dynamic treatment rules. But there are a few issues needed to clarify. For example, are black-box methods better or worse than interpretable methods like tree-like method? In this talk, I will discuss a few issues about reinforcement learning in estimating dynamic treatment rules.

Identify Sensitive Biomarkers of Alzheimer's Disease with Longitudinal Block-Wise Missing Data using Multiple Imputations Across Different Sources
♦Zhongzhe Ouyang and Lu Wang
University of Michigan, Ann Arbor
The Alzheimer's Disease Neuroimaging Initiative (ADNI) study is a precision health initiative designed to develop effective treatments that can slow or stop the progression of Alzheimer's Disease (AD). Since there is no uniform method for early diagnosis of AD, personalized treatment strongly relies on the identification of sensitive biomarkers. In this paper, we aim to locate these biomarkers using the ADNI dataset. We propose a multiple imputation method to handle block-wise missing covariates in ADNI longitudinal data, where missing covariates are imputed based on all covariates from complete data source and partial covariates from observed sources including complete data source. We construct estimating equations with imputed data for each source and then aggregate the information across sources with generalized method of moments. We adopt SCAD penalty in the variable selection and use EBIC criteria for tuning parameter selection. We derived the consistency, sparsity, and normality of the proposed estimator, and demonstrate the outperformance of the proposed method with numerical experiments.

Session 23INT97: Recent Developments on the Analysis of Censored Data

On Interquantile Smoothness of Censored Quantile Regression with Induced Smoothing (Cqris)
♦Zexi Cai[1] and Tony Sit[2]
[1]Columbia University
[2]The Chinese University of Hong Kong

Quantile regression has emerged as a useful and effective tool in modeling survival data, especially for cases where noises demonstrate heterogeneity. Despite recent advancements, non-smooth components involved in censored quantile regression estimators may often yield numerically unstable results, which, in turn, lead to potentially self-contradicting conclusions. We propose an estimating equation-based approach to obtain consistent estimators of the regression coefficients of interest via the induced smoothing technique to circumvent the difficulty. Our proposed estimator can be shown to be asymptotically equivalent to its original unsmoothed version whose consistency and asymptotic normality can be readily established. Extensions to handle functional covariate data and recurrent event data are also discussed. To alleviate the heavy computational burden of bootstrap-based variance estimation, we also propose an efficient resampling procedure that reduces the computational time considerably. Our numerical studies demonstrate that our proposed estimator provides substantially smoother model parameter estimates across different quantile levels and can achieve better statistical efficiency compared to a plain estimator under various finite-sample settings. The proposed method is also illustrated via four survival datasets, including the HMO HIV data, the primary biliary cirrhosis (PBC) data, and so forth.

Distributed Censored Quantile Regression

⬥*Tony Sit[1] and Kelly Xing[2]*
[1]CUHK
[2]Michigan State University

This talk discusses an extension of censored quantile regression to a distributed setting. With the growing availability of massive datasets, it is oftentimes an arduous task to analyse all the data with limited computational facilities efficiently. Our proposed method, which attempts to overcome this challenge, consists of two key steps, namely: (i) estimation of both Kaplan-Meier estimator and model coefficients in a parallel computing environment; (ii) aggregation of coefficient estimations from individual machines. We study the upper limit of the order of the number of machines for this computing environment, which, if fulfilled, guarantees that the proposed estimator converges at a comparable rate to that of the oracle estimator. In addition, we also provide two further modifications for distributed systems including (i) a divide-and-conquer approximation in the sense of Chen et al. (2019) and (ii) a nonparametric counterpart along the direction of Kong & Xia (2017) for censored quantile regression. Numerical experiments are conducted to compare the proposed and the existing estimators. The promising results demonstrate the computation efficiency of the proposed methods. Finally, for practical concerns, a cross validation procedure is also developed which can better select the hyperparameters for the proposed methodologies.

Semiparametric Regression Analysis of Doubly-Censored Data with Applications to Incubation Period Estimation

⬥*Kin Yau Wong[1], Qingning Zhou[2] and Tao Hu[3]*
[1]The Hong Kong Polytechnic University
[2]The University of North Carolina at Charlotte
[3]Capital Normal University

The incubation period is a key characteristic of an infectious disease. In the outbreak of a novel infectious disease, accurate evaluation of the incubation period distribution is critical for designing effective prevention and control measures. Estimation of the incubation period distribution based on limited informa-tion from retrospective inspection of infected cases is highly challenging due to censoring and truncation. In this paper, we consider a semiparametric regression model for the incubation period and propose a sieve maximum likelihood approach for estimation based on the symptom onset time, travel history, and basic demographics of reported cases. The approach properly accounts for the pandemic growth and selection bias in data collection. We also develop an efficient computation method and establish the asymptotic properties of the proposed estimators. We demonstrate the feasibility and advantages of the proposed methods through extensive simulation studies and provide an application to a dataset on the outbreak of COVID-19.

A Semiparametric Joint Model for Cluster Size and Sub-unit-specific Interval-censored Outcomes

⬥*Chun Yin Lee[1], Kin Yau Wong[1], Kwok Fai Lam[2] and Dipankar Bandyopadhyay[3]*
[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong
[2]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, Hong Kong
[3]Department of Biostatistics, Virginia Commonwealth University, Virginia, USA

Clustered data frequently arise in biomedical studies, where observations, or subunits, measured within a cluster are associated. The cluster size is said to be informative, if the outcome variable is associated with the number of subunits in a cluster. In most existing work, the informative cluster size issue is handled by marginal approaches based on within-cluster resampling, or cluster-weighted generalized estimating equations. Although these approaches yield consistent estimation of the marginal models, they do not allow estimation of within-cluster associations and are generally inefficient. We propose a semiparametric joint model for clustered interval-censored event time data with informative cluster size. We use a random effect to account for the association among event times of the same cluster as well as the association between event times and the cluster size. For estimation, we propose a sieve maximum likelihood approach and devise a computationally-efficient expectation-maximization algorithm for implementation. The estimators are shown to be strongly consistent, with the Euclidean components being asymptotically normal and achieving semiparametric efficiency. Extensive simulation studies are conducted to evaluate the finite-sample performance, efficiency and robustness of the proposed method.

Session 23INT99: Recent Development of Tensor Time Series

Optimal Subsampling Bootstrap for Massive Data

⬥*Yingying Ma[1], Chenlei Leng[2] and Hansheng Wang[3]*
[1]Beihang University
[2]University of Warwick
[3]Peking University

The bootstrap is a widely used procedure for statistical inference because of its simplicity and attractive statistical properties. However, the vanilla version of bootstrap is no longer feasible computationally for many modern massive datasets due to the need to repeatedly resample the entire data. Therefore, several improvements to the bootstrap method have been made in

recent years, which assess the quality of estimators by subsampling the full dataset before resampling the subsamples. Naturally, the performance of these modern subsampling methods is influenced by tuning parameters such as the size of subsamples, the number of subsamples, and the number of resamples per subsample. In this paper, we develop a novel hyperparameter selection methodology for selecting these tuning parameters. Formulated as an optimization problem to find the optimal value of some measure of accuracy of an estimator subject to computational cost, our framework provides closed-form solutions for the optimal hyperparameter values for subsampled bootstrap, subsampled double bootstrap and bag of little bootstraps, at no or little extra time cost. Using the mean square errors as a proxy of the accuracy measure, we apply our methodology to study, compare and improve the performance of these modern versions of bootstrap developed for massive data through numerical study. The results are promising.

Huber Principal Component Analysis for Large-Dimensional Factor Model

♦Yong He[1], Lingxiao Li[1], Dong Liu[2] and Wen-Xin Zhou[3]
[1]Shandong University
[2]Shanghai University of Finance and Economics
[3]University of California, San Diego

Factor models have been widely used in economics and finance. However, the heavy-tailed nature of macroeconomic and financial data is often neglected in the existing literature. To address this issue and achieve robustness, we propose an approach to estimate factor loadings and scores by minimizing the Huber loss function, which is motivated by the equivalence of conventional Principal Component Analysis (PCA) and the constrained least squares method in the factor model. We provide two algorithms that use different penalty forms. The first algorithm, which we refer to as Huber PCA, minimizes the ℓ_2-norm-type Huber loss and performs PCA on the weighted sample covariance matrix. The second algorithm involves an element-wise type Huber loss minimization, which can be solved by an iterative Huber regression algorithm. Our study examines the theoretical minimizer of the element-wise Huber loss function and demonstrates that it has the same convergence rate as conventional PCA when the idiosyncratic errors have bounded second moments. We also derive their asymptotic distributions under mild conditions. Moreover, we suggest a consistent model selection criterion that relies on rank minimization to estimate the number of factors robustly. We showcase the benefits of Huber PCA through extensive numerical experiments and a real financial portfolio selection example. An R package named "HDRFA"(https://cran.r-project.org/web/packages/HDRFA/index.html) has been developed to implement the proposed robust factor analysis.

Online Change-Point Detection for Matrix-Valued Time Series with Latent Two-Way Factor Structure

Yong He, Xinbing Kong, Lorenzo Trapani and ♦Long Yu
This paper proposes a novel methodology for the online detection of changepoints in the factor structure of large matrix time series. Our approach is based on the well-known fact that, in the presence of a changepoint, the number of spiked eigenvalues in the second moment matrix of the data increases (e.g., in the presence of a change in the loadings, or if a new factor emerges). Based on this, we propose two families of procedures - one based on the fluctuations of partial sums, and one based on extreme value theory - to monitor whether the first non-spiked eigen-

value diverges after a point in time in the monitoring horizon, thereby indicating the presence of a changepoint. Our procedure is based only on rates; at each point in time, we randomise the estimated eigenvalue, thus obtaining a normally distributed sequence which is i.i.d. with mean zero under the null of no break, whereas it diverges to positive infinity in the presence of a changepoint. We base our monitoring procedures on such sequence. Extensive simula- tion studies and empirical analysis justify the theory. An R package implementing the procedure is available on CRAN.

Session 23INT102: Statistical Inference on Complex/Compositional Data and Biostatistics

Fdr Control for Linear log-Contrast Models with High-Dimensional Compositional Covariates

♦Gaorong Li, Panxu Yuan and Changhan Jin
Beijing Normal University

Linear log-contrast models have been widely used to describe the relationship between the response and the compositional covariates, in which one central task is to identify the significant compositional covariates while achieving false discovery rate (FDR) control. This paper proposes an FDR control method for linear log-contrast models with high-dimensional compositional covariates by completely bypassing traditional p-values. Under some regularity conditions, we provide theoretical guarantees for the proposed method in terms of FDR control, and the theoretical power is also proven to approach one as the sample size tends to infinity. The finite-sample performance of the proposed method is evaluated through extensive simulation studies, and applications to microbiome compositional datasets are also provided.

Renewable Quantile Regression with Heterogeneous Streaming Datasets

♦Xuerong Chen and Senlin Yuan

The renewable statistical inference have received much attention since the advent of streaming data collection techniques. However, most existing online updating methods are developed based on a homogeneity assumption and gradients; all data batches are required to be either independent and identically distributed or share the same regression parameters, and objective functions must be smooth concerning parameters. To our best knowledge, the only existing approach that allows some regression parameters to be different for different data batches, was proposed by Luo and Song (2021) who required the homogeneous structure to be known, which is difficult to guarantee in actual application. In this paper, we develop an online renewable quantile regression method that relies only on the current data and summary statistics of historical data, for both homogeneous and heterogeneous streaming data. The proposed methods are computationally efficient, can automatically detect the unknown potential homogeneous structure, and are robust to heavy-tailed noise and data with outliers. Asymptotic properties show that the proposed renewable estimators can achieve the same statistical efficiency as the oracle estimators based on individual level data. A numerical simulation and a real data analysis illustrate that the proposed methods perform well.

Augmented Two-Step Estimating Equations with Nuisance

Functionals and Complex Survey Data

♦*Puying Zhao*[1] *and Changbao Wu*[2]
[1]Yunnan University
[2]University of Waterloo

Statistical inference in the presence of nuisance functionals with complex survey data is an important topic in social and economic studies. The Gini index, Lorenz curves and quantile shares are among the commonly encountered examples. The nuisance functionals are usually handled by a plug-in nonparametric estimator and the main inferential procedure can be carried out through a two-step generalized empirical likelihood method. Unfortunately, the resulting inference is not efficient and the nonparametric version of the Wilks' theorem breaks down even under simple random sampling. We propose an augmented estimating equations method with nuisance functionals and complex surveys. The second-step augmented estimating functions obey the Neyman orthogonality condition and automatically handle the impact of the first-step plug-in estimator, and the resulting estimator of the main parameters of interest is invariant to the first step method. More importantly, the generalized empirical likelihood based Wilks' theorem holds for the main parameters of interest under the design-based framework for commonly used survey designs, and the maximum generalized empirical likelihood estimators achieve the semiparametric efficiency bound. Performances of the proposed methods are demonstrated through simulation studies and an application using the dataset from the New York City Social Indicators Survey.

Session 23INT29: Recent Advances on Interplay of Statistics and Optimization

Self-Regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models

Yury Polyanskiy[1] *and* ♦*Yihong Wu*[2]
[1]MIT
[2]Yale

Title: Self-regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models Abstract: Introduced by Kiefer and Wolfowitz 1956, the nonparametric maximum likelihood estimator (NPMLE) is a widely used methodology for learning mixture models and empirical Bayes estimation. Sidestepping the non-convexity in mixture likelihood, the NPMLE estimates the mixing distribution by maximizing the total likelihood over the space of probability measures, which can be viewed as an extreme form of overparameterization. In this work we discover a surprising property of the NPMLE solution. Consider, for example, a Gaussian mixture model on the real line with a subgaussian mixing distribution. Leveraging complex-analytic techniques, we show that with high probability the NPMLE based on a sample of size n has $O(\log n)$ atoms (mass points), significantly improving the deterministic upper bound of n due to Lindsay 1983. Notably, any such Gaussian mixture is statistically indistinguishable from a finite one with $O(\log n)$ components (and this is tight for certain mixtures). Thus, absent any explicit form of model selection, NPMLE automatically chooses the right model complexity, a property we term self-regularization. Statistical applications and extensions to other exponential families will be given. Time permitting, we will discuss recent progress

and open problems on the optimal regret in empirical Bayes and the role of NPMLE. This is based on joint work with Yury Polyanskiy (MIT): https://arxiv.org/abs/2008.08244 and https://arxiv.org/abs/2109.03943

Random Graph Matching at Otter's Threshold via Counting Chandeliers

Cheng Mao[1]*, Yihong Wu*[2]*,* ♦*Jiaming Xu*[3] *and Sophie Yu*[3]
[1]Georgia Institute of Technology
[2]Yale University
[3]Duke University

We propose an efficient algorithm for graph matching based on similarity scores constructed from counting a certain family of weighted trees rooted at each vertex. For two ER graphs whose edges are correlated through a latent vertex correspondence, we show that this algorithm correctly matches all but a vanishing fraction of the vertices with high probability, provided that the average degree diverges and the edge correlation coefficient squared is above Otter's tree-counting constant. Moreover, this almost exact matching can be made exact under an extra condition that is information-theoretically necessary. This is the first polynomial-time graph matching algorithm that succeeds at an explicit constant correlation and applies to both sparse and dense graphs. In comparison, previous methods either require the correlation to converge to 1 or are restricted to sparse graphs. The crux of the algorithm is a carefully curated family of rooted trees called chandeliers, which allows effective extraction of the graph correlation from the counts of the same tree while suppressing the undesirable correlation between those of different trees. Based on joint work with Cheng Mao (Gatech), Yihong Wu (Yale), and Sophie H. Yu (Duke). Preprint available at https://arxiv.org/pdf/2209.12313.pdf

Snr Estimation under High-Dimensional Linear Models

Xiaohan Hu and ♦*Xiaodong Li*
UC Davis

Estimation of signal-to-noise ratios and residual variances in high-dimensional linear models has important applications including heritability estimation in bioinformatics. Random effects likelihood estimators have been widely used in practice for SNR estimation, and it is known to be consistent when the model is misspecified. In this talk, we aim to investigate the conditions on both the design matrix and the coefficient vector, such that asymptotic behaviors for this SNR estimator can be explicitly derived. We will stress tools from random matrix theory and normal approximation of quadratic forms. For future work, extensions to method-of-moments, diverging aspect ratios, and linear models with feature groups will be briefly discussed. This is a joint work with my student Xiaohan Hu.

Session 23INT104: Innovative Designs and Analysis Methods for Clinical Trials and Complex Data

Least Squares Support Vector Regression for Complex Censored Data

Xinrui Liu, Xiaogang Dong, Le Zhang, Jia Chen and ♦*Chunjie Wang*
Changchun University of Technology

Least squares support vector regression (LS-SVR) is a robust machine learning algorithm for small sample data. Its solution is derived from solving a set of linear equations, making the

calculation process straightforward. In order to overcome the difficulties of the regression estimations when the responses are subject to interval censoring or left truncation and right censoring, two LS-SVR methods are proposed. For interval-censored data, one can easily estimate the regression functions by combining the imputation techniques and LS-SVR for right-censored data. For left-truncated and right-censored data, a weight is used to reduce the effects of truncation and censoring on the LS-SVR procedure. Simulation results show that the proposed methods can reduce regression error and yield high accuracy and stability.

Estimating Optimal Individual Treatment Regimes in Semi-Supervised Framework

♦*Mengjiao Peng and Yong Zhou*

Finding the optimal individualized treatment rule mapping from the individual characteris- tics or contextual information to the treatment assignment has been studied intensively in the literature, with important applications in practice. We consider the problem of estimating the optimal treatment regime in a semi-supervised learning setting, where a very small proportion of the entire set of observations are labeled with the true outcome but features predictive of the outcome are available among all observations. We propose a model-free robust inference approach for optimal treatment regime by the aid of the unlabeled data with only covariate in- formation to improve estimation efficiency. The proposed estimation of OPT primarily involves a flexible nonparametric imputation by single index kernel smoothing which works well even for high-dimensional covariates; and a follow-up estimation for optimal treatment regime based on concordance-assisted learning, including optimization of the estimated concordance function up to a threshold and finding the optimal threshold to maximize the inverse propensity score weighted (IPSW) estimator of the value function. Moreover, when the propensity score function is unknown, a doubly robust estimation method is developed under a class of monotonic index models. Our estimators are shown to be consistent and asymptotically normal. Simulations exhibit the efficiency and robustness of the proposed method compared to existing approaches in finite samples.

Session 23INT59: Recent Advances in Biomedical Data Science

Do We Need the Target-Decoy Strategy in Peptide Identification?

Sheng Lian[1], Juntao Zhao[1], Zhen Zhang[2], Xiaodan Fan[2], Ning Li[1] and ♦*Weichuan Yu[1]*

[1]HKUST
[2]CUHK

The target-decoy strategy has been widely used in peptide identification to estimate the false discovery rate (FDR). But the construction of the decoy database has been criticized since such an approach was introduced in 2007 for not being able to guarantee the "randomness" in the decoy database. In this talk, we like to tackle this issue from the hypothesis testing point of view. We argue that the decoy database may not be necessary when calibrating the confidence of large-scale peptide identification using mass spectrometry.

Functional Adaptive Double-Sparsity Estimator for Func-

tional Linear Regression with Applications in Kinect Sensor Analysis and Automated Elderly Health Assessments

Xinyue Li
City University of Hong Kong

Kinect sensors have been increasingly used in healthcare to provide mobility assessments for the elderly and monitor patients performing intervention or rehabilitation exercises. A Kinect sensor system contains cameras to monitor a person's body joint movements at high frequency, yielding a skeletal dataset that contains twenty-five joint coordinate time series in the x, y, z directions. A Kinect device can collect the person's joint movement trajectories over time as a person performs functional activities and captures how the person conducts postural control. Therefore, Kinect sensors coupled with mobility assessment tools have the potential to examine which joints and how joints play a crucial role in postural control and develop an automatic assessment tool. However, how to analyze high-frequency multi-dimensional Kinect data is challenging. This talk will discuss our proposed method to effectively analyze Kinect sensor data and study the association with mobility evaluation outcomes. We treat time-series data from each joint in each axis as a function, and thus Kinect sensors yield multi-dimensional functional data. To fully leverage the information contained in Kinect coordinate time-series data, we further obtained the velocity and acceleration data by calculating derivatives and obtained frequency data by applying the Fast Fourier Transform. With multi-dimensional functional Kinect data, we further applied penalization approaches to identify which joint in which dimension is associated with mobility assessment outcomes. Our penalization approach first achieves global sparsity to select which joint function is helpful for assessing mobility. Then it further achieves local sparsity to identify which part of the joint function is associated with mobility assessment outcomes. We term the combination of global sparsity and local sparsity as double sparsity. We applied our proposed method to a Kinect sensor dataset collected during mobility assessments in Hong Kong elderly community centers to identify the detailed association between joint movements and elderly mobility and functionality. Furthermore, we developed an automatic mobility assessment tool for the elderly, which saves human labor, allows for continuous self-assessments and immediate feedback in the elderly community centers, and enables early detection of balance and gait deficits and timely intervention to enhance elderly healthcare.

A Multi-Use Graph Neural Network Framework for Single-Cell Multi-Omics Data

♦*Peifeng Ruan[1] and Hongyu Zhao[2]*

[1]UT Southwestern Medical Center
[2]Yale University

The advances of single-cell multi-omics profiling technologies in biomedical research offer an unprecedent opportunity for understanding cell heterogeneity and subpopulations. There are many statistical and computational challenges in the integrative analyses of these rich data, including sequencing sparsity, complex differential patterns in gene expression, and different platforms and panels used to generate multiple single-cell multi-omics. In this presentation, we introduce a multi-use graph neural network framework that can effectively impute and predict missing sequencing panels, integrate multi-omics single-cell datasets, and formulate and aggregate cell–cell relationships with graph neural

networks. Comprehensive simulations and applications on multiple CITE-seq and single-cell RNA-sequencing datasets demonstrate that our proposed method is a powerful tool for general single-cell data multi-omics analyses that outperforms the existing methods for protein prediction, gene imputation and cell clustering.

Session 23INT105: Recent Developments on Variable Selection and Regression Analysis with Censored Data

Debiased Average Distance Correlation Screening of Massive Interval Censored Data under Orthogonal Subsampling

Huiqiong Li
Yunnan University
TBA

Bel and Bayesian Variables Selection for Partially Linear Models Based on Right-Censored Data

♦*Chunjing Li*[1] *and Xiaogang Dong*
[1]lichunjing@ccut.edu.cn
This paper, we consider the Bayesian empirical likelihood (BEL) method and the variable selection for partially linear models with right-censored data. We extended BEL to the semiparametric regression model under right-censored data. By consructing the BEL confidence interval of the model parameters and corresponding algorithm, the asymptotic posterior distribution of BEL can be obtained. The BEL method has obvious advantages over the EL method under the right-censored data semiparametric regression model. It avoids the complicated calculation of variance. Then based on the spike-and-slab prior, we proposed the Bayesian variable selection of partially linear models under the censored data using the Bayesian hierarchical model. In the simulation study, the effect of popular algorithms and Bayesian variable selection algorithms was demonstrated, which Bayesian variable selection has higher accuracy and correct identification rate in a limited sample. In the empirical analysis, the right-censored data of acute Myelogenous leukemia was verified, and the empirical results are very significant.

Simultaneous Variable Selection and Estimation for Joint Models of Longitudinal and Failure Time Data with Interval Censoring

Fengting Yi
Yunnan University
This paper discusses variable selection in the context of joint analysis of lon- gitudinal data and failure time data. A large literature has been developed for either variable selection or the joint analysis but there exists only limited litera- ture for variable selection in the context of the joint analysis when failure time data are right censored. Corresponding to this, we will consider the situation where instead of right-censored data, one observes interval-censored failure time data, a more general and commonly occurring form of failure time data. For the problem, a class of penalized likelihood-based procedures will be developed for simultaneous variable selection and estimation of relevant covariate effects for both longitudinal and failure time variables of interest. In particular, a Monte Carlo EM (MCEM) algorithm is presented for the implementation of the pro- posed approach. The proposed method allows for the number of covariates to be diverging with the sample size and is shown to have the oracle property. An extensive simulation study is con- ducted to assess the finite sample performance of the proposed approach and indicates that it works well in practical situations. An application is also provided.

Semiparametric Regression Analysis of Length Biased Interval-Censored Data under the Mixture Cured Model

♦*Peijie Wang, Cunjin Zhao and Jianguo Sun*
Jilin University
Interval-censored failure time data frequently occur in many areas and a large literature on their analyses has been established. In this paper, we discuss the situation where one faces length biased interval-censored data, which are commonly encountered in prospective follow-up studies. For the problem, the nonparametric maximum likelihood estimation for the proportional mixture cured model is developed and a computationally simple and stable EM algorithm is given. The resulting estimators of regression parameters are shown to be consistent and asymptotically normal. A simulation is conducted to assess the finite sample performance of the proposed method and suggests that it performs well. Finally the proposed method is applied to the real data.

Session 23INT106: Analyzing Big and Complex Data using Modern Machine Learning Techniques

Fragmgan: Generative Adversarial Nets for Fragmentary Data Imputation and Prediction

♦*Fang Fang and Shenliao Bao*
East China Normal University
Modern scientific research and applications very often encounter "fragmentary data"which brings big challenges to imputation and prediction. By leveraging the structure of response patterns, we propose a unified and flexible framework based on Generative Adversarial Nets (GAN) to deal with fragmentary data imputation and label prediction at the same time. Unlike most of the other generative model based imputation methods that either have no theoretical guarantee or only consider Missing Completed At Random (MCAR), the proposed FragmGAN has theoretical guarantees for imputation with data Missing At Random (MAR) while no hint mechanism is needed. FragmGAN trains a predictor with the generator and discriminator simultaneously. This linkage mechanism shows significant advantages for predictive performances in extensive experiments.

Statistical Methods for Allele-Specific Expression Analysis using Single-Cell Rna-Seq Data

Rui Xiao
University of Pennsylvania
Allele-specific gene expression (ASE) analysis, an alternative and complementary approach to eQTL analysis, is a powerful tool for identifying variation in gene expression. ASE quantifies the relative expression of two alleles in a diploid individual, and the imbalance of expression of the two alleles may explain phenotypic variation and disease pathophysiology. ASE is driven by cis-regulatory variants located near a gene. Since the two alleles used to measure ASE are expressed in the same cellular environment and genetic background, they can serve as internal controls and eliminate the influence of trans-acting genetic and environmental factors. In this talk, I will focus on statistical methods for ASE analysis using RNA sequencing (RNA-seq) and single-cell RNA-seq (scRNA-seq) data that we recently de-

veloped. Specifically, I will first introduce a statistical model for detection of gene-level ASE across multiple individuals in a population under one clinical condition, as well as ASE difference between two clinical conditions. ASE patterns may vary across cell types. To better identify cellular targets of disease, we will next introduce a recently developed statistical method to characterize cell-type-specific ASE in bulk RNA-seq data by incorporating cell type composition information inferred from external scRNA-seq data. This method is extended to detect genes whose cell-type-specific ASE are associated with clinical factors by modeling covariate effect.

Robust Multiple Testing under High Dimensioanl Factor Model

Xinxin Yang[1] and ♦Lilun Du[2]

[1]INNO Asset Management

[2]City University of Hong Kong

Large-scale multiple testing under static factor models is commonly used to select skilled funds in financial market. However, static factor models are arguably too stringent as it ignores the serial correlation, which severely distorts error rate control in large-scale inference. In this manuscript, we propose a new multiple testing procedure under dynamic factor models that is robust against both heavy-tailed distributions and the serial dependence. The idea is to integrate a new sample-splitting strategy based on chronological order and a two-pass Fama-Macbeth regression to form a series of statistics with marginal symmetry properties and then to utilize the symmetry properties to obtain a data-driven threshold. We show that our procedure is able to control the false discovery rate (FDR) asymptotically under high-dimensional dynamic factor models. As a byproduct that is of independent interest, we establish a new exponential-type deviation inequality for the sum of random variables on a variety of functionals of linear and non-linear processes. Numerical results including a case study on hedge fund selection demonstrate the advantage of the proposed method over several state-of-the-art methods.

Learning Individualized Minimal Clinically Important Difference (Imcid) from High-Dimensional Data

Jiwei Zhao

University of Wisconsin Madison

Statistical significance has been widely used to infer the treatment effect in assessing the efficacy of a treatment or intervention; however, there has been a growing recognition that statistical significance has its own limitations. Clinical significance, on the contrary, is usually desirable in practice as it provides a better assessment of the clinically meaningful improvement. A critical concept in evaluating clinical significance is minimal clinically important difference (MCID), the smallest change in the outcome that an individual patient would identify as important. In this talk, I will present a statistical learning framework for estimating the individualized MCID (iMCID) from high-dimensional data. In particular, I will present a path-following iterative algorithm and some novel nonregular theoretical results. Additionally, simulation studies that reinforce our theoretical findings and an application to the study of chondral lesions in knee surgery to demonstrate the usefulness of the proposed approach will also be discussed.

Session 23INT108: High-Dimensional Statistical Inference

Nonparametric Granger Causality Test Based on Adaptive Dimension Reducing Technique

♦Yongchang Hui, Hong Jiang and Yi Liu

Xi'an Jiaotong University
TBC

Change Point Inference in the High-Dimensional Correlation Matrix

♦Zhaoyuan Li and Jie Gao

CUHK-SZ

This paper considers the problem of detecting a change point and estimating the location in the correlation matrices of a sequence of high-dimensional vectors, where the dimension is large enough to be comparable to the sample size or even much larger. A new break test is proposed based on signflip parallel analysis to detect the existence of a change point. Furthermore, a two-stage approach combining a signflip permutation dimension reduction step and a CUSUM statistic is proposed to estimate the change point's location and recover the support of changes. Our method also successes in estimating the change point in the tail by combing the SMOTE method. Theoretical and numerical properties are investigated to support the new methodologies.

Separable Sample Covariance Matrices under Elliptical Populations with Applications

Huiqin Li[1], Guangming Pan[2], ♦Yanqing Yin[1] and Wang Zhou[3]

[1]Chongqing University

[2]Nanyang Technological University

[3]National University of Singapore

We investigate the spectral properties of separable covariance matrices under elliptical populations. The separable covariance matrix model can handle both cross-row and cross-column correlations thus gain more popularity recently. Under the high-dimensional setting where the dimension p and the sample size n tend to infinity proportionally, we find the limit of the empirical spectral distribution and establish the central limit theorems (CLT) for linear spectral statistics (LSS) of such kinds of sample covariance matrices. Some applications of our established CLT are also given.

Adaptive Tests for Bandedness of High-Dimensional Covariance Matrices

Xiaoyi Wang

Beijing Normal University

Estimation of high-dimensional banded covariance matrix is widely used in multivariate statistical analysis. To ensure the validity of estimation, we aim to test the hypothesis that the covariance matrix is banded with a certain bandwidth under the high-dimensional framework. Though several testing methods have been proposed in the literature, the existing tests are only powerful for some alternatives with certain sparsity levels, whereas they may not be powerful for alternatives with other sparsity structures. The goal of this paper is to propose a new test for the banded covariance matrix, which is powerful for alternatives with various sparsity levels. The proposed new test can also be used for testing the banded structure of covariance matrices of error vectors in high-dimensional factor models. Simulation studies and an application to a prostate

cancer dataset from protein mass spectroscopy are conducted for evaluating the validity of the proposed new tests for the banded covariance matrix.

Index of Authors

www.icsa.org