The 11th ICSA International Conference

Hangzhou, China
December 20-22, 2019

International Chinese Statistical Association
泛華統計協會

International Chinese Statistical Association

The 11th ICSA International Conference

CONFERENCE INFORMATION AND PROGRAM

December 20-22, 2019
Hangzhou Dragon Hotel
Hangzhou, China

Organized by
International Chinese Statistical Association
Co-organized by
Center for Data Science, Zhejiang University

Contents

Welcome

The 11th ICSA International Conference

December 19-22, Hangzhou Dragon Hotel
Hangzhou, China

Welcome to the 11th International Chinese Statistical Association (ICSA) International Conference!

The 11th ICSA International Conference is held at Hangzhou Dragon Hotel, Hangzhou, China, on December 19-22, 2019. The theme of this conference is: **Innovation with Statistics and Data Science**.

The organizing and program committees have been working diligently to put together a strong and comprehensive program to provide a wealth of activities and opportunities for discussion and exchange. The conference program consists of 170 sessions, including three special lectures, as well as exciting social events. The special lectures will be given by three distinguished statisticians: Peter Hall Lecture by ***Dr. Jianqing Fan*** (*Princeton University*), Pao-Lu Hsu Award Lecture by ***Dr. Hongyu Zhao*** (*Yale University School of Public Health*), *and* Keynote Lecture by ***Dr. Zhiliang Ying*** (*Columbia University*). The conference highlights methodological and applied contributions of statistics, mathematics, and computer science in modern data science. It brings together the statistical community and scientists from related fields to present, discuss and disseminate research and best practices.

This conference attracts more than 700 statisticians working in academia, government, and industry from all over the world, including Mainland China, Hong Kong, Taiwan, Japan, South Korea, India, Singapore, Australia, New Zealand, United States, Canada, Chile, United Kingdom, Switzerland, Belgium, Sweden and Israel. We hope the conference offers you great opportunities for learning, networking and recruiting, and that you will receive inspiration from the presented research to develop new ideas. Social events in the 11th ICSA International Conference include a banquet (Friday, December 20 evening). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

Hangzhou is the capital city of Zhejiang Province and serves as a local political, economic and cultural center. As the southern terminus of the Grand Canal, the city is located on the lower reaches of the Qiantang River in southeast China. The subtropical monsoon climate contributes to varied seasonal sceneries, making Hangzhou one of China's most popular travel destinations all year-round. The West Lake is undoubtedly the most renowned landmark, noted for its scenic beauty that blends naturally with many famous historical and cultural sites. The "Ten West Lake Prospects," selected from the most frequently visited attractions around the lake, gives travelers a panoramic view of the city's highlights. Take a stroll along the causeway by the lake; you'll feel the peaceful ethos of the city and better understand its time-honored fame as 'Heaven on Earth'.

Thank you for coming to the 11th ICSA International Conference in Hangzhou!

Hongzhe Li, Ph.D., Chair, The 11th ICSA International Conference Program Committee

Intruduction of Center for Data Science

Center for Data Science, Zhejiang University

“浙江大学数据科学研究中心”成立于 2017 年 5 月 18 日，是以统计学、应用数学、计算机科学和管理学为核心支撑学科，以大数据理论、应用研究和人才培养为主的校级学术创新研究机构。研究中心强调与经济学、医学、生命科学、社会学、工学等众多学科领域的交叉融合，在获取基础研究的硕果同时，注重于科研技术成果的转化。面向市场和产业需求进行人才培养，着重培养研究型和技术型人才，同时通过组建跨学科、跨领域的教学、研究师资队伍，大力培养复合型数据科学人才。中心发展的方向包括研究数据科学的理论与方法；为处理海量的、高维的数据发展分析的工具和算法，进而为应用和管理提供方向；设立数据科学的教学平台，培养相关领域内的国际化高端人才。研究中心致力于面向大数据战略前瞻问题和社会需求，为不同学科之间实现有渠道的、有组织的合作提供平台，凝聚和培养数据科学一流人才，打造高质量、高效率的数据科学核心技术与理论研究创新平台。研究中心基本定位：汇聚和培养数据科学高端人才，开展大数据原创性研究与开发应用，切实解决中国大数据领域的前瞻性核心问题。研究中心的核心目标：通过机制创新，汇聚和培养国内外数据科学领域的一流人才，构建数据科学高精尖科研创新基地。利用杭州大数据行业的优势和浙江大学多学科交叉的特点，尽快使浙江大学成为国际大数据研究领域的引领者之一。

The Center for Data Science, Zhejiang University was established on May 18th, 2017. The Center is a university-level, cross-disciplinary platform which is based on statistics, applied mathematics, computer science and management. The Center is aimed to embark on seminal research in the analysis of complex data, including statistical theory and applications, machine learning, and algorithm design, to train advanced professionals in data science, and to serve the government, society and industry in all aspects.

Intruduction of Center for Data Science

Recruitment

Center for Data Science, Zhejiang University

Zhejiang University is a top research-orientated university located in Hangzhou, which is one of the most beautiful cities in China. The university has strong ties with Hangzhou's dynamic and fast-growing community of high-tech and internet companies. With generous funding from Zhejiang University, The Center for Data Science was established in May 2017, which is aimed to embark on seminal research in the analysis of complex data, including statistical theory and applications, machine learning, and algorithm design, to train advanced professionals in data science, and to serve the government, society and industry in all aspects.
The Center invites applications for Assistant/Associate/Full Professor faculty positions. Applicants with doctoral degrees in statistics, computer science, and other areas related to the analysis of big data and its applications are encouraged to apply.

• Position Title
Full-time Assistant /Associate/Full Professor

• Duties and Responsibilities
Candidates are expected to conduct independent and innovative research, cooperate in scientific research projects, and to teach courses at all levels.

• Position Qualifications
Candidates should demonstrate the potential for excellence in teaching and research, and are expected to publish papers in top journals. Preferences will be given to those with teaching and research experience in statistical theory and methodology, data mining, big data analytics, econometrics, and machine learning.

• Benefits
1. Salary Package (including basic salary, relocation fund, and other benefits) will be competitive and commensurate with qualifications. In addition, the Center supports high-quality research with generous research grants.
2. Housing is provided according to university policy.

Applicants are required to email the following application materials to stazlx@zju.edu.cn with the subject line "Application for faculty positions to the Center for Data Science":
(1) Cover letter
(2) CV
(3) 3 representative research papers (full text)
(4) Evidence of teaching performance (if available).
Please also arrange 3 reference letters to be sent to stazlx@zju.edu.cn.

• Contact Email
Professor ZHANG
E-mail: stazlx@zju.edu.cn

ICSA COMMITTEES

ICSA 2019 MEMBERS OF THE COMMITTEES

Program Committee

Chair:

- Hongzhe Li，University of Pennsylvania

Committee members:

- Tianxi Cai，Harvard University
- Minghui Chen，University of Conneticut
- Song Xi Chen，Peking University
- Jeng-Min Chiou，Academia Sinica
- Aurore Delaigle， University of Melbourne
- Ke Deng，University of Melbourne
- Peng Ding，University of California, Berkeley
- Wing Kam Fung，Hong Kong University
- Zhi Geng，Peking University
- Anil Ghosh， Indian Statistical Institute
- Tony Guo，Beigene
- Zijian Guo，Rutgers University
- Fang Han，University of Washington
- Ruth Heller，Tel Aviv University
- Shimodaira Hidetoshi，Kyoto University
- Hu, Jianhua Hu，Columbia University
- Hongkai Ji，Johns Hopkins University
- Jiashun Jin，Carnege-Mellon University
- Zhezhen Jin，Columbia University
- Mei-ling Lee，University of Maryland
- Chenlei Leng， Universityof Warwick
- Yingying Li，Hongkong University of Science and Technology
- Huazhen Lin，Southwestern University of Finance and Economics
- Xiwu Lin，Janssen Research
- Weidong Liu，Shanghai Jiaotong University
- Wendy Lou，University of Toronto
- Henry Lu，National Chiao Tung University, Taiwan
- Wenbin Lu，North Carolina State University
- Ying Lu，Stanford University
- Shuangge Ma，Yale University
- Zhaoling Meng，Gates Foundation
- Hans Mueller，University of California, Davis
- Bin Nan，University of Califonia, Irvine
- Guangming Pan，Nanyang Technology University
- James Pan，Johnson and Johnson
- Byeong Park，Seoul National University
- Limin Peng，Emory University
- Long Qi，University of Pennsylvania
- Annie Qu，University of Illinois
- Hui Quan，Sanofi
- Wei Shen，Eli Lilly

- Jianxin Shi，National Cancer Institute
- Haochang Shou，University of Pennsylvania
- Liuquan Sun，Institute of Applied Mathematics, CAS
- Tony Sun，University of Missouri
- Wei Fred Sun，Hutchison Cancer Research Center
- Niansheng Tang，Yunnan University
- Junhui Wang，City University of Hong Kong
- Mei Cheng Wang，Johns Hopkins University
- Wanjie Wang，National University of Singapore
- Xueqin Wang，Sun Yat-Sen University
- Yazhen Wang，University of Wisconsin
- Hulin Wu，University of Texas, Houston
- Yin Xia，Fudan University
- Ronghui Xu，University of California, San Diego
- Qiwei Yao，The London School of of Economics
- Bingming Yi，Vertex
- Grace Yi，University of Waterloo
- Jianxin Yin，Renming University of China
- Ming Yuan，Columbia University
- Anru Zhang，University of Wisconsin
- Lixin Zhang，Zhejiang University
- Lu Zhang，Google
- Shurong Zheng，Northeastern Normal University
- Qin Zhou，University of California, Los Angeles
- Hongtu Zhu，DiDi Chuxing
- Ji Zhu，University of Michigan
- Changliang Zou，Nankai University
- Fei Zou，University of North Carolina, Chapel Hill

ICSA COMMITTEES

New Researcher Award Committee

Chair：
Annie Qu，University of Illinois

Committee members:
Ji Zhu，University of Michigan
Bin Nan，University of California，Irvine
Ronghui Xu，University of California，San Diego
Wei Sun，Fred Hutch
Weidong Liu，Shanghai Jiaotong University
Junhui Wang，City University of Hongkong
Ke Deng，Tsinghua University

Organizing Committee

Chairs：
Gang Li, Janssen R&D
Lixin Zhang, Zhejiang University
Hongzhe Li, University of Pennsylvania

Committee members:

Dong Han, Shanghai Jiaotong University
Rui Feng, University of Pennsylvania
Hangjin Jiang, Zhejiang University
Hongzhe Li, Zhejiang University
Junhong Lin, Zhejiang University
Wei Luo, Zhejiang University
Xiaoye Miao, Zhejiang University
Zhonggen Su, Zhejiang University
Renjun Xu, Zhejiang University
Rongmao Zhang, Zhejiang University

Fund raising committee:

Gang Li, Fund Raising Chair, Janssen R&D
Hongzhe Li, Fund Raising Co-chair, University of Pennsylvania
Zhonggen Su, Zhejiang University

Transportation

Location

The 11th ICSA International Conference will be held at Hangzhou Dragon Hotel. The address is 120 Shuguang Road, Xihu District, Hangzhou, China.

Fly to Hangzhou Xiaoshan International Airport (HGH) :

Airport Bus: Airport Bus Wulin Gate Line (Wulin Gate Civil Aviation Ticket Office Direction)
Airport - Wulin Gate Civil Aviation Ticket Office Station (transfer No. 28 (or No. 92, 318b, 318)
Wulin Gate Station - Songmuchang Station, 100 meters (about 2 minutes) walk to Hangzhou Dragon Hotel.
Taxi: about 125 yuan (+20 yuan road toll)

Take the high-speed rail to Hangzhou East Railway Station:

Bus: 28 Hangzhou East Railway Station West (Botanical Garden Direction) - Pine Wood Yard Station 100 meters (about 2 minutes) walk to Hangzhou Dragon Hotel
Taxi: about 30 yuan

Take the high-speed rail to Hangzhou Station:

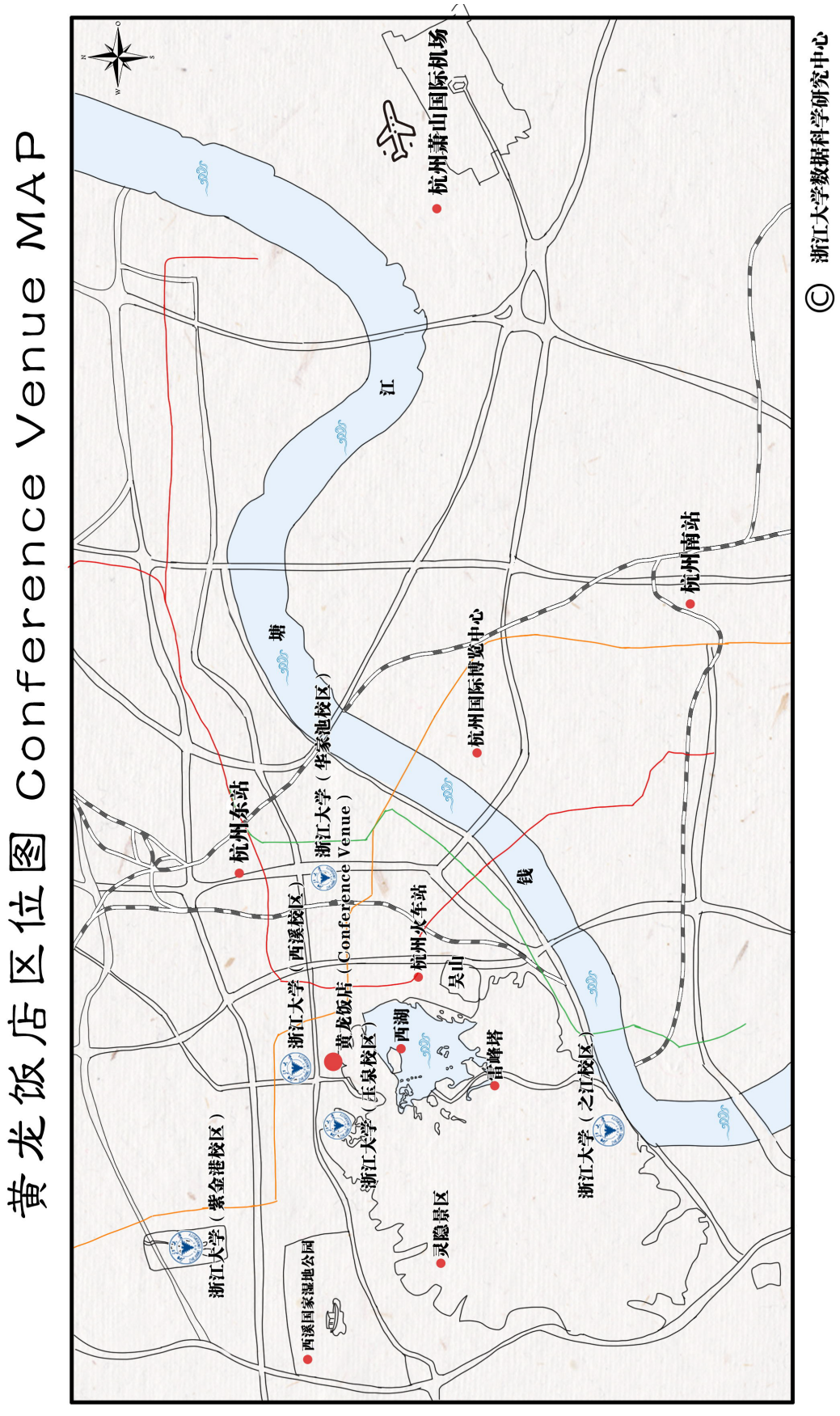
Bus: 49 Lucheng Station Railway Station (Bus West Station Direction) - Pine Wood Yard Station, 100 meters (about 2 minutes) walk to Hangzhou Dragon Hotel.
Taxi: about 20 yuan

Transportation

Shuttle Bus

Date	Departure Time	Departing from	Stop	Destination
Dec 20, Friday	8:00	Yuquan Hotel	Union Lingfeng Hotel	Hangzhou Dragon Hotel
Dec 21, Saturday	8:30	Yuquan Hotel	Union Lingfeng Hotel	Hangzhou Dragon Hotel
Dec 22, Sunday	8:30	Yuquan Hotel	Union Lingfeng Hotel	Hangzhou Dragon Hotel

Venue Map

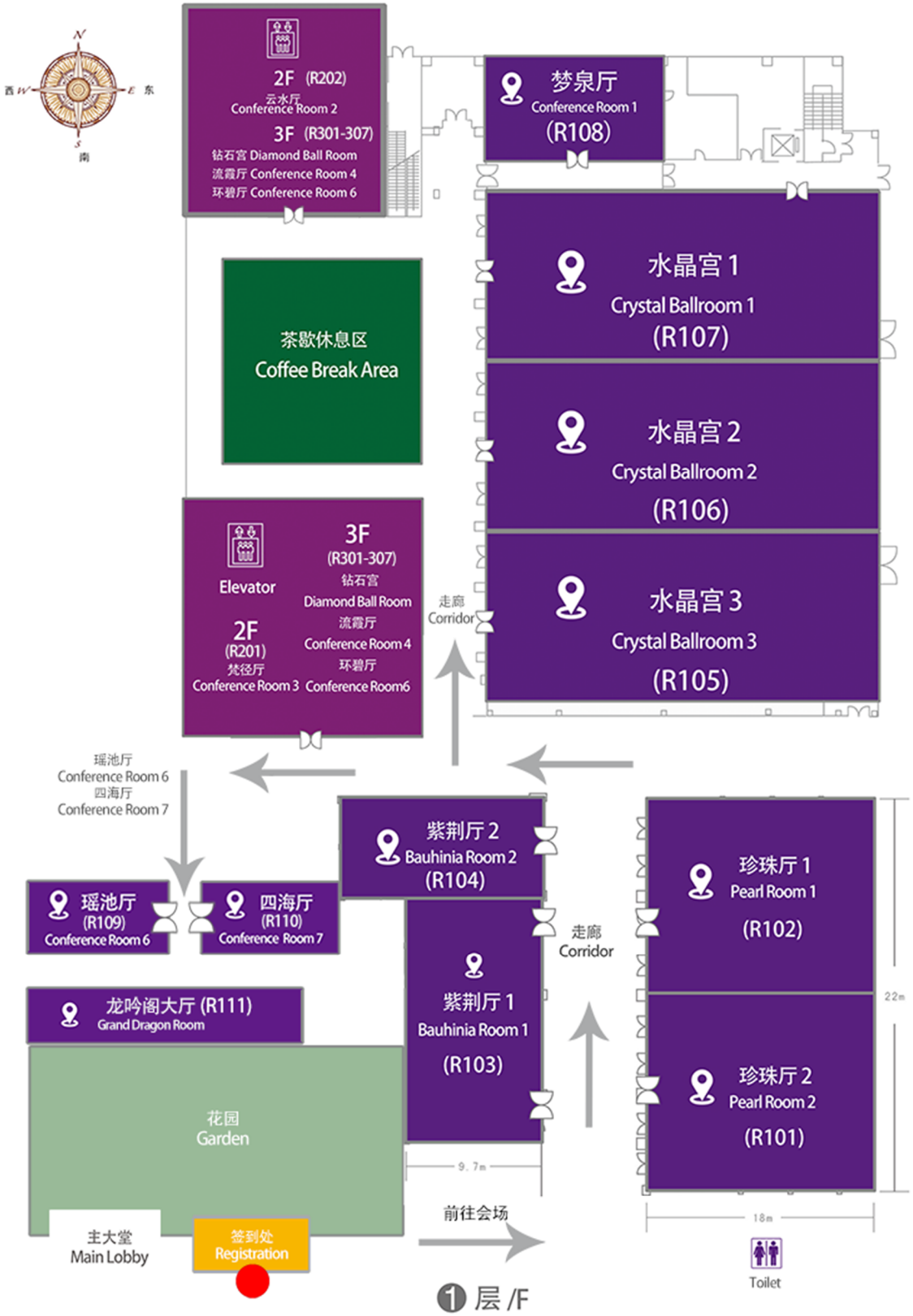

黄龙饭店区位图 Conference Venue MAP
杭州萧山国际机场
江
塘
杭州东站
浙江大学（华家池校区）
浙江大学（西溪校区）
黄龙饭店（Conference Venue）
杭州国际博览中心
杭州南站
杭州火车站
吴山
西湖
雷峰塔
钱
浙江大学（紫金港校区）
浙江大学（玉泉校区）
浙江大学（之江校区）
西溪国家湿地公园
灵隐景区
© 浙江大学数据科学研究中心

Venue Map

ICSA2019 黄龙酒店会场分布图 总览

Venue Map

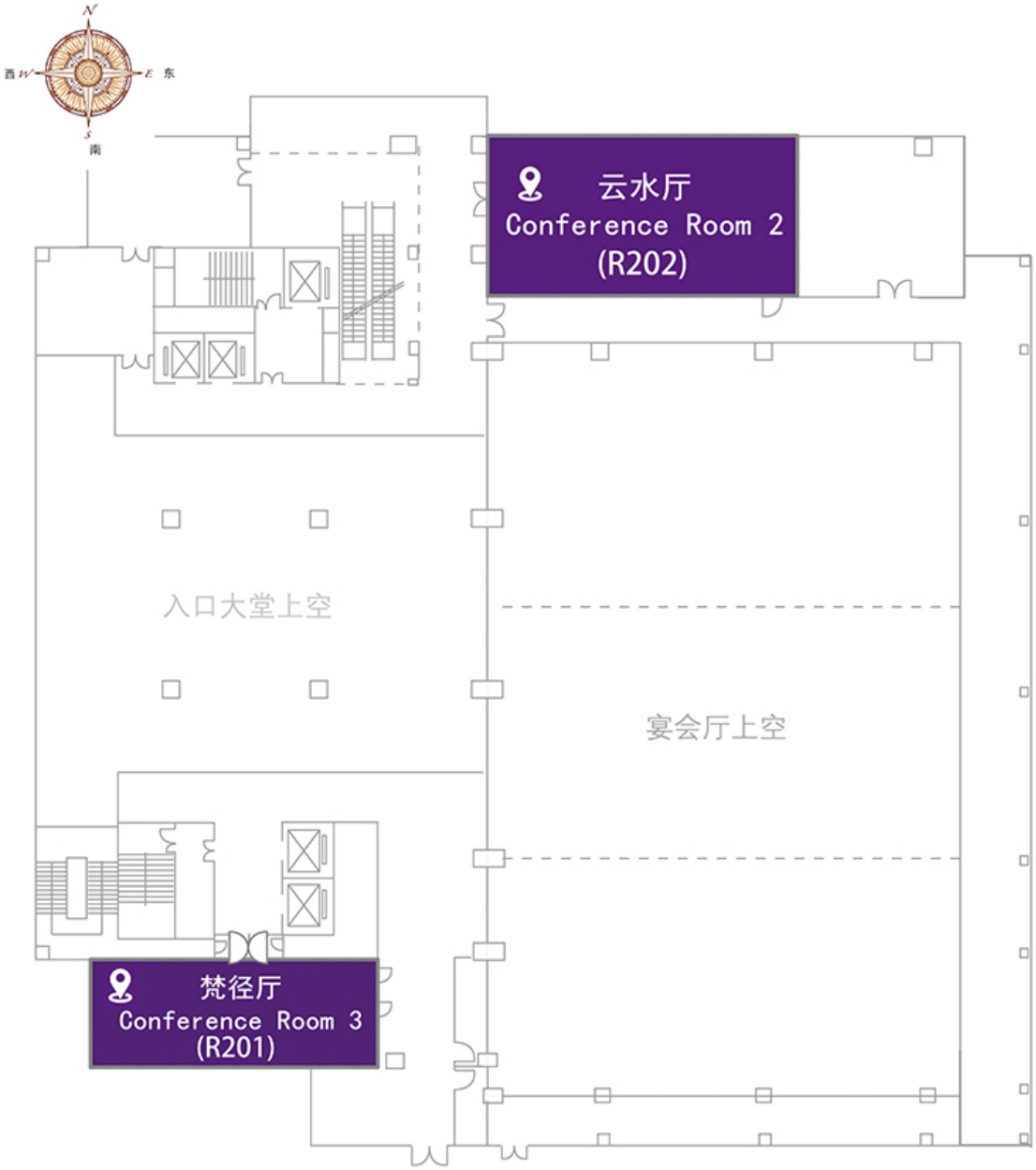
Venue Map



ICSA2019 会场分布图

Venue Map

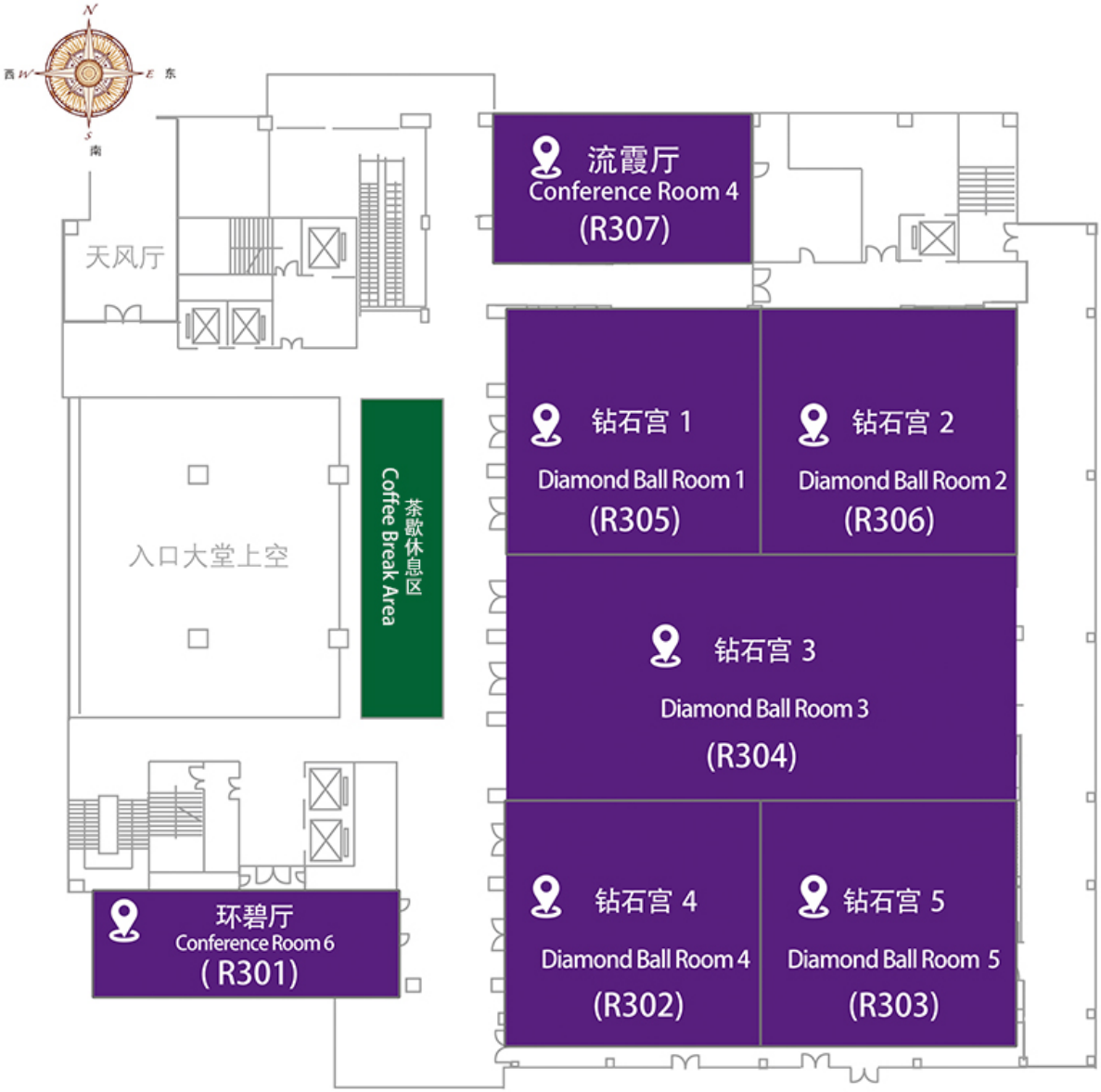
Venue Map



❷层/F

ICSA2019会场分布图

Venue Map



❸ 层/F

ICSA2019 会场分布图

Venue Map

Dining and Wi-Fi Information

Downtown Dining Information

1、Huanglong business circle：Huanglong business circle is located in huanglong distribution center, 1.2 kilometers away from huanglong hotel. There are a lot of delicious food，like A chicken hotpot，Old uncle snack，Starbucks Coffee，Heyi Japanese cuisine...
黄龙商圈：黄龙商圈位于黄龙集散中心，距离黄龙饭店 1.2 公里。这里有很多的美食，比如一席地鸡窝，老娘舅快餐，星巴克咖啡，和一料理...

2、Lakeside business circle：Lakeside business circle is 2.5 kilometers away from huanglong hotel. It is mainly featured by tourism, leisure and high-grade shopping. Qingchun road in the north, connected with wulin business circle, jiefang road in the south, zhongshan middle road in the east, and lakeside road in the west. Hangzhou is one of the well-known business district. Of course, a lot of food is gathered here. Huanglong business circle is located in the center of huanglong distribution, 1.2 kilometers away from huanglong hotel. There are a lot of delicious food，like Ginger hudong barbecue，Grandma's Home，KFC,Xinyuyuan Creative hangzhou cuisine...
湖滨商圈：湖滨商圈距离黄龙饭店 2.5 公里，以旅游休闲、高档购物为主要消费特色。北至庆春路，与武林商圈相连，南到解放路，东邻中山中路，西沿湖滨路。是杭州知名的商圈之一。当然这里也聚集了很多的美食，比如姜虎东烤肉，外婆家,肯德基，新榆园创意杭帮菜...

3、Wulin business circle：Wulin business circle is located in the center of hangzhou city, 2 kilometers away from huanglong hotel. This has always been the most prosperous core business district of hangzhou, and there are numerous delicacies here. Like De shou gong South Korean dish，Ge lao guan Bullfrog hot pot，McDonald's，Carbon Italian restaurant...
武林商圈：武林商圈位居杭州城市中心，距离黄龙饭店 2 公里。这里一直是杭州最繁华的核心商圈，这里的美食也是数不胜数，有德寿宫韩国料理，哥老关重庆美蛙鱼头，麦当劳，莫卡哆意大利餐厅...

4、Qingzhi wood：Qingzhi wood is located in the north of yugu road, between zhejiang university yuquan campus and botanical garden, 3 kilometers away from huanglong hotel. At the entrance stands a landscape stone with the words "green zhi wu" in big characters. After the stone, is a wang bibo ripples artificial lake - "qingliu pond", the lake has pro level platform, there are curved pavilion gallery. Here, also gathered a lot of food，Like Pu Shu Hangzhou cuisines，Nan shan nan art-themed restaurant，Huai gu izakaya，Firefox...
青芝坞：青芝坞位于玉古路北侧，浙大玉泉校区和植物园之间，距离黄龙饭店 3 公里。入口处矗立着一块写的“青芝坞”三个大字的景观石。石后，是一汪碧波荡漾的人工湖——“青柳塘”，湖边有亲水平台，有弯曲的亭廊。在这里，也聚集了很多的美食，有朴墅传承杭帮菜，南山南文艺主题餐厅，怀古居酒屋，火狐狸...

Wi-Fi Information

Kind reminder for your WIFI connection：

Guest Room：Select “The Dragon” network. Open the browser then click “login”.

Public Area：Select “The Dragon” network. Open the browser, use your 5-digit room number as both the user ID and password.

Banquet

Location and Time: Hangzhou Dragon Hotel Crystal Ballroom, December 20 (Friday), 6:30 pm-9:00 pm (fee event, ticket required).

Lunch Breaks

Box-Lunch Provided on Dec 20, 21, 22, which is included in the registration fee.
Time: Dec 20/21/22 12:10-13:30
Rooms :

1F	R101 R102 R103 R104
3F	R302 R303 R304 R305 R306

Program Overview

Program Overview

Friday December 20, 2019

Time	Room	Session
8:30 am-9:00 am	Crystal Ballroom	Opening Ceremony
9:00 am-10:00 am	Crystal Ballroom	Peter Hall Lecture
Coffee Break		
10:30 am-12:10 am	*See Program*	Parallel Sessions
Lunch Break		
1:30 pm-3:10 pm	*See Program*	Parallel Sessions
Coffee Break		
3:40 pm-5:20 pm	*See Program*	Parallel Sessions
6:30 pm-9:00 pm	Crystal Ballroom	Banquet

Saturday December 21, 2019

Time	Room	Session
9:00 am-9:10 am	Crystal Ballroom	Award Ceremony
9:10 am-10:00 am	Crystal Ballroom	Pao-Lu Hsu Lecture
Coffee Break		
10:30 am-12:10 am	*See Program*	Parallel Sessions
Lunch Break		
1:30 pm-3:10 pm	*See Program*	Parallel Sessions
Coffee Break		
3:40 pm-5:20 pm	*See Program*	Parallel Sessions

Sunday December 22, 2019

Time	Room	Session
9:00 am-10:00 am	Crystal Ballroom	Keynote Lecture
Coffee Break		
10:30 am-12:10 am	*See Program*	Parallel Sessions
Lunch Break		
1:30 pm-3:10 pm	*See Program*	Parallel Sessions
Coffee Break		
3:40 pm-5:20 pm	*See Program*	Parallel Sessions

Peter Hall Lecture

Jianqing Fan, Princeton University

Jianqing Fan, Ph,D. is a statistician and financial econometrician. He is Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and former Chairman of Department of Operations Research and Financial Engineering at the Princeton University. He is the winner of the 2000 COPSS Presidents' Award, Morningside Gold Medal for Applied Mathematics (2007), and Guggenheim Fellow (2009), Pao-Lu Hsu Prize (2013) and Guy Medal in Silver (2014). He was elected to Academia Sinica in 2012.

Title: Communication-Efficient Accurate Statistical Estimation

Location and Time：Crystal Ballroom, December 20 (Friday), 9:00 am-10:00 am

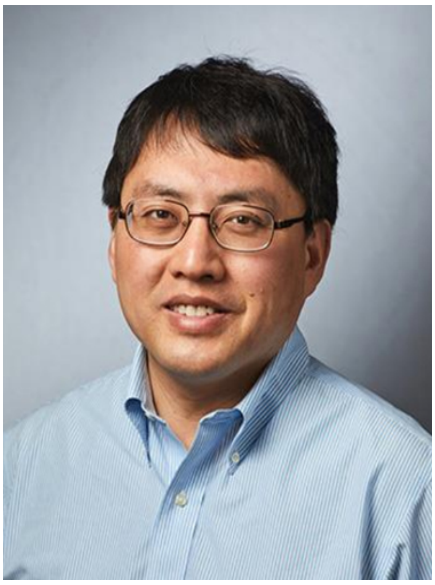
Chair: T. Tony Cai, University of Pennsylvania

Abstract:

When the data are stored in a distributed manner, direct application of traditional statistical inference procedures is often prohibitive due to communication cost and privacy concerns. This paper develops and investigates two Communication-Efficient Accurate Statistical Estimators (CEASE), implemented through iterative algorithms for distributed optimization. In each iteration, node machines carry out computation in parallel and communicates with the central processor, which then broadcasts aggregated gradient vector to node machines for new updates. The algorithms adapt to the similarity among loss functions on node machines, and converge rapidly when each node machine has large enough sample size. Moreover, they do not require good initialization and enjoy linear converge guarantees under general conditions. The contraction rate of optimization errors is derived explicitly, with dependence on the local sample size unveiled. In addition, the improved statistical accuracy per iteration is derived. By regarding the proposed method as a multi-step statistical estimator, we show that statistical efficiency can be achieved in finite steps in typical statistical applications. In addition, we give the conditions under which one-step CEASE estimator is statistically efficient. Extensive numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the superior performance of our algorithms.

(Joint work with Yongyi Guo and Kaizheng Wang)

Pao-Li Hsu Award Lecturee



Hongyu Zhao, Yale School of Public Health

Hongyu Zhao, Ph.D. is the Ira V. Hiscock Professor of Biostatistics and Professor of Statistics and Data Science and Genetics, Chair of the Biostatistics Department and the Co-Director of Graduate Studies of the Inter-Departmental Program in Computational Biology and Bioinformatics at Yale University. His research interests are the applications of statistical methods in genetics, molecular biology, drug developments, and precision medicine.

Dr. Zhao is a Co-Editor of the Journal of the American Statistical Association Theory and Methods, and serves on the editorial boards of several leading statistical and genetics journals. He was the recipient of the Mortimer Spiegelman Award for a top statistician in health statistics under the age of 40 awarded by the American Public Health Association and the Pao-Lu Hsu Award from the International Chinese Statistical Association. His research has also been recognized by the Evelyn Fix Memorial Medal and Citation by UC Berkeley, a Basil O'Connor Starter Scholar Award by the March of Dimes Foundation, election to the fellowship of the American Association for the Advancement of Science, the American Statistical Association, and the Institute of Mathematical Statistics.

Title: Fisher's 1918 Quantitative Genetics Model In the Genomics Era
Location and Time：Crystal Ballroom, December 21 (Saturday), 9:00 am-10:00 am
Chair: Heping Zhang, Yale University

Abstract:
In 1918, R. A. Fisher reported a comprehensive study of a statistical model relating an individual's quantitative traits to his/her genetic factors in his seminal paper entitled "The Correlation between Relatives on the Supposition of Mendelian Inheritance". This model laid the foundation for the field of quantitative genetics. More than a century later, this model still proves effective in understanding the genetic basis of human complex traits when tens of thousands of chromosomal regions have been implicated for hundreds of traits through Genome-Wide Association Studies (GWAS) in the past 15 years. In this presentation, I will discuss how Fisher's model has been used to quantify the genetic contributions to complex traits using GWAS results, its robustness to model misspecifications, and its extensions to identify relevant tissues/cell types for a specific trait and genetic correlations between different traits. I will also discuss statistical inference using either individual genotype and phenotype data, a typical set up for traditional statistical analysis, or summary statistics, which are more easily accessible for GWAS data. This is joint work with Can Yang, Jiming Jiang, Qiongshi Lu, Debashis Paul, Wei Jiang, Cecilia Dao, Yiliang Zhang, and others.

Zhiliang Ying, Columbia University

Zhiliang Ying, Ph.D. has been a professor with Department of Statistics at Columbia University since 2000. Zhiliang Ying's research areas include survival analysis, semiparametric models, sequential analysis, educational and psychological measurement, latent variable models, stochastic control, statistical methods for financial data among other. He has published over 150 journal articles and supervised 40 PhD students.

Zhiliang Ying is an elected fellow of the Institute of Mathematical Statistics (1995) and of the American Statistical Association (1999). He served as President of the International Chinese Statistics Association (ICSA). He is a recipient of the Morningside Gold Medal of Applied Mathematics (2004), ICSA Distinguished Achievement Award (2007), National Council on Measurement in Education (NCME) Annual Award (2008), AERA Signicant Contribution to Educational Measurement and Research Methodology Award (2011) and ICSA Outstanding Service Award (2018).

Zhiliang Ying has served as Co-Editor of Statistica Sinica and as associate editor for many journals: Annals of Statistics, Biometrics, Journal of American Statistical Association, Scandinavian Journal of Statistics, Bernoulli among others. He has also served on the National Science Foundation panels and the National Institutes of Health study sections.

Title: Statistical models and methods for educational and psychological measurement

Location and Time：Crystal Ballroom, December 22 (Sunday), 9:00 am-10:00 am

Chair: Hongzhe Li, University of Pennsylvania

Abstract:

Statistical models have played important and fundamental roles in educational and psychological measurement. The increased computing power and data collection capability provide new opportunities as well as challenges. The first part of this talk covers the classical item response theory models which have been widely used in standardized testing, as well as recent developments on related multidimensional latent factor/class models with focus on important issues such as local independence or lack of it, identifiability among others. The second part covers modeling and analysis of process data arise from modern computer-based tests with items for assessing complex problem solving skills in technology-rich environments. Examples from educational assessment and psychological evaluation will be used throughout the presentation.

New Researcher Awards

New Researcher Awards

Fei Xue, University of Illinois

Title: Integrating multi-source block-wise missing data in model selection
Time: Dec 21 (Saturday) 15:40-17:20
Room: R105
S009: Theoretical challenges for estimations and predictions for large-scale data

Zijian Guo, Rutgers University

Title: Inference for Case Probability in High-dimensional Logistic Regression
Time: Dec 21 (Saturday) 10:30-12:10
Room: R104
S007：Recent Advances on Large Complex Data

Xin Zhang, Florida State University

Title: Subspace-regularized Tensorial Parameter Estimation
Time: Dec 20 (Friday) 10:30-12:10
Room: R305
S030：Statistical Inference for High-dimensional Tensor Data

Anru Zhang, University of Wisconsin-Madison

Title: High-dimensional Tensor Regression Analysis
Time: Dec 20 (Friday) 10:30-12:10
Room: R305
S030：Statistical Inference for High-dimensional Tensor Data

Timothy Cannings, University of Edinburgh

Title: Classification with imperfect training labels
Time: Dec 22 (Sunday) 15:40-17:20
Room: R101
S015：Regression and classification for complex data

Yang Ni, Texas A&M University

Title: Covariate-dependent graphs with application in cancer genomics
Time: Dec 21 (Saturday) 15:40-17:20
Room: R201
S156：Recent advances in complex biological data modeling

Jue Hou, Harvard T.H. Chan School of Public Health

Title: Estimating Treatment Effect under Additive Hazards Models with High-dimensional Covariates
Time: Dec 22 (Sunday) 13:30-15:10
Room: R304
S038：Treatment Effects and Other Emerging Issues in Biomedical Data Science

Fan Zhou, Shanghai University of Finance and Economics

Title: Statistics, Optimization and Deep Learning in the ride-sharing Industry

Time: Dec 20 (Friday) 10:30-12:10
Room: R307
S109：Statistical and Machine Learning Methods with Application in AI Transportation

Masaaki Imaizumi, The Institute of Statistical Mathematics
Title: Generalization Analysis for Mechanism of Deep Learning via Nonparametric Statistics
Time: Dec 22 (Sunday) 10:30-12:10
Room: R305
S105：Statistical Theory for Neural Networks and Machine Learning

Tianxi Li, University of Virginia
Title: Hierarchical community detection by recursive partitioning
Time: Dec 20 (Friday) 15:40-18:30
Room: R304
S159：Recent Develpments in Statistical Network Analysis

Anderson Zhang, The Wharton School, University of Pennsylvania
Title: Optimality of Spectral Clustering for Gaussian Mixture Model
Time: Dec 21 (Saturday) 10:30-12:10
Room: R102
S055：Random Matrices Theory and Applications

Sponsors and Career Services

The 11th ICSA International Conference Sponsors

The 11th ICSA International Conference Sponsors Committees gratefully acknowledge the generous support of our sponsors below.

Gold Sponsors

BeiGene

DiDi

用正医药 YONGZHENG CRO

燃石医学 Burning Rock Dx

Silver Sponsors

NEKTAR®

先声药业 Simcere

The ICSA 2019 sponsors currently have the following statistics related openings.

BeiGene

BeiGene, Ltd. ("BeiGene") is a commercial-stage biotechnology company focused on developing and commercializing innovative molecularly-targeted and immuno-oncology drugs for the treatment of cancer.

BeiGene has a broad portfolio consisting of six internally-developed, clinical-stage, drug candidates, including three late-stage clinical drug candidates, zanubrutinib (BTK inhibitor), tislelizumab (PD-1 antibody), and pamiparib (PARP inhibitor). BeiGene has also in-licensed six drugs and drug candidates, including three marketed drugs in China ABRAXANE®, REVLIMID® and VIDAZA® under an exclusive license from Celgene Corporation, and two clinical-stage drug candidates with development and commercialization rights in China and other selected countries in the Asia-Pacific region.

BeiGene was established in Beijing in 2010 and listed on the U.S. NASDAQ Global Select Market in February 2016. As of October 16, 2019, the Company had a global team of over 3,000 employees. BeiGene operates as a fully-integrated global biotechnology company with broad capabilities, both in China and globally, spanning research, clinical development, manufacturing and commercialization. The Statistics and Data Science department of BeiGene has four functions including Biostatistics, Scientific programming, Data Management, System and Standard. With over 300 staffs based in six offices in Beijing, Shanghai, Wuhan, San Francisco, New Jersey and Boston, the group provides analytic support for BeiGene's business activities including research and translational science, clinical development and post approval activities.

Sponsors and Career Services

About DiDi Chuxing

Didi Chuxing ("DiDi") is the world's leading mobile transportation platform. The company offers a full range of app-based transportation options for 550 million users, including Taxi, Express, Premier, Luxe, Bus, Designated Driving, Enterprise Solutions, Bike Sharing, E-bike Sharing, Car Rental and Sharing and Food Delivery. DiDi is committed to work with communities and partners to solve the world's transportation, environmental and employment challenges using big data-driven deep-learning algorithms. By continuously improving user experience and creating social value, DiDi strives to build an open, efficient, and sustainable transportation ecosystem.

DiDi is hiring!

Positions: Data Scientist, Data Analysis Expert, Business Analyst

Ideal candidates should:

- Be passionate about transportation industry and excited to make real-world impact. We are moving people in physical world.
- Be rigorous in analytical practice, but versatile in various business contexts.
- Be able to extract business insights from massive data, and turn insights into business proposal and actions.
- Be self-driven, and be able to prioritize projects to achieve highest impact.
- Have quantitative background (math, stats, econ, operation research, computer science, etc.) or relevant experience. Be hands on (SQL, R/Python, Tableau, etc.).

Contact: Mandy Ma, Research Outreach Manager mandyma.edu@didiglobal.com

DiDi

用正医药
YONGZHENG CRO

Company Profile

Shanghai Yongzheng Medical Science and Technology Co., Ltd. is one of the earliest clinical contract research organizations (CRO) established in China in 2004, focusing on providing full clinical services for pharmaceutical products research and development. Yongzheng is a CRO with 15 years of steady and sound running state. The headquarter is in Shanghai with 2 subsidiaries, "Bujin" and "Helpclin Data", which provides SMO and data management & biostatistics services respectively. Yongzheng covers more than 80% regions in China. Yongzheng has service outlets including but not limited to Shanghai, Beijing, Wuxi, Tianjin, Harbin, Changchun, Shenyang, Guangzhou, Nanning, Chengdu, Wuhan, Changsha, Zhengzhou, Nanchang, Xi 'an, etc.).

There are over 300 employees which all have medical or pharmaceutical background. 80% of the core members have worked in multinational companies, received professional training from pharmaceutical companies and CRO, and have rich experience in clinical research of drugs and medical devices. Yongzheng has successively taken hundreds of clinical trials including Phase I~IV and post market trials. Yongzheng has experiences in collaborating with many clinical professional departments and therapeutic areas.

Yongzheng concentrates on developing its core business such as trial management, data management & biostatistics and pharmacovigilance. Simultaneously, Yongzheng also develops other business which related to the core business, such as regulatory affairs, medical affairs and SMO. Yongzheng has been implementing the service concept of "Quality Oriented, Efficiency Prioritized, Innovation Driven". We focus on project management and consolidation of service foundation, which forms Yongzheng's service advantage.

公司简介

上海用正医药科技有限公司成立于 2004 年，是国内最早成立的临床合同研究组织（CRO）之一，专注于为医药产品研发提供全方位临床服务。用正医药总部位于上海，历经 15 年的稳健发展，旗下设立"北京和普润"、"成都步锦"两家全资子公司，分别提供数据管理与生物统计、SMO 服务。用正医药覆盖全国 80%以上地区，在上海、北京、无锡、天津、哈尔滨、长春、沈阳、广州、南宁、成都等国内主要城市设立了服务网点。

超过 300 人的专业技术团队均具备医学或药学背景，80%核心成员曾供职过跨国企业，在药物、器械临床研究方面经验丰富。用正医药先后承担了数百个Ⅰ~Ⅳ期及上市后再评价临床试验项目，十多年的行业浸润成就了用正医药在多个治疗领域的丰富经验。

用正医药集中资源发展试验管理、数据管理与生物统计、药物警戒等主营业务，同时也发展与主营业务相关的注册和医学事务、SMO 等业务，是一家可以提供全方位服务的临床 CRO。用正医药多年来秉承"质量先导•效率优先•创新驱动"的服务理念，专注项目管理，夯实服务基础，形成用正服务优势。

Sponsors and Career Services

燃石医学
Burning Rock Dx

我们是谁：

LinkedIn 评选的 Top25 创业公司。燃石医学成立于 2014 年，专注于为肿瘤精准医疗提供最具临床价值的二代基因测序（NGS）服务。目前，燃石医学已经针对不同癌种和临床场景开发了 30 余种检测产品，涉及肿瘤靶向和免疫用药伴随诊断、肿瘤良恶性鉴别、微小残留病灶监测、肿瘤复发进展预测和肿瘤敏感性检测。

燃石医学拥有中国第一间由美国 CLIA 认证的 ctDNA 和肿瘤组织 NGS 实验室，并同时获得美国 CAP 实验室认证和国家卫生健康委临床检验中心（National Center for Clinical Laboratories, NCCL）“高通量测序实验室”技术审核。。公司与全国 400 多家顶尖医院开展了深度合作，并积累了中国最大的实体瘤基因组数据库之一。

除了提供 LDT 模式（临床实验室自建项目）的 NGS 检测服务，燃石医学也是一站式 NGS 平台解决方案、推进高质量 NGS 检测落地医院病理科从而惠及更多肿瘤患者的先行者。公司拥有首个获 NMPA（中国国家药品监督管理局）批准的 NGS 检测试剂盒，与安捷伦科技、 珀金埃尔默、Illumina、凯杰以及众多制药公司达成战略性合作，将始终致力于开发创新与可靠的 NGS 检测产品，为肿瘤患者带来临床获益。

招聘职位：生信科学家（研发部）

- 熟悉生信算法，计算机，统计学，应用数学等相关专业背景博士
- 熟悉一种或多种编程语言（C++/java/python/perl/R/matlab）
- 希望你有扎实的统计学基础如数理统计、概率论和数学基础比如抽象代数，微分方程等
- 熟悉 pattern recognition，signal deconvolution 等高纬数据分析建模
- 熟悉 Linux 开发环境
- 了解 CNV 等流程检测的基本原理
- 良好的抽象和逻辑分析能力，以及对问题本质的提炼

招聘职位：生信工程师-算法（研发部）

- 生物信息学，统计学，应用数学/工程等相关专业硕士
- 熟练掌握一种或多种编程语言（C++/java/python/perl/R/matlab）
- 具备统计学基础知识，对概率论，数理统计有一定基础
- 对数据降维、特征选择、机器学习等算法开发有课题经验者优先
- 熟悉 Linux 开发环境，有全基因组/全外显子（WGS/WES），全甲基化组（WGBS），或全转录组(RNA-seq)等二代测序分析经验优先
- 有创新性研究经验者优先考虑
- 良好的抽象和逻辑分析能力

招聘职位：生信工程师/生信科学家（生物信息部）

- 生物学，生物信息学，计算机，统计学，应用数学等相关专业背景博士或硕士学位均可
- 具有 NGS 分析基础，了解典型生信算法的基本原理，有流程（整体流程或某个子功能）维护经验者优先
- 熟悉 Linux 开发环境，熟悉 bash 语言， 擅长 R 或其他一种或多种编程语言（python/perl/matlab）
- 有一定的数据分析经验，具有统计学或机器学习背景者优先
- 良好的抽象和逻辑分析能力，以及对问题本质的提炼

欢迎您投递简历至：rui.zhang@brbiotech.com

ICSA 2020 Announcement

ICSA 2020 Applied Statistics Symposium (May 17-20, 2020)

The ICSA Applied Statistics Symposium will be held from Sunday, May 17 to Wednesday, May 20, 2020 at The Westin Galleria Houston, 5060 W Alabama Street, Houston, Texas. Please send any inquiry to Dr. Hulin Wu (Hulin.Wu@uth.tmc.edu). Please visit https://symposium2020.icsa.org/ for details including the key dates.

Call for Invited Session Proposals

The symposium scientific program committee welcomes invited session proposals. An invited session consists of either 4 presenters or 3 presenters and 1 discussant. The one-talk rule will be applied (i.e., each speaker can only give one invited talk). It is required to confirm all speakers' availability before proposal submission. Of particular interest are sessions that appeal to diverse audiences and are closely related to the symposium theme Advancing Statistics for Data Intelligence. You may increase the chance of the proposal to be accepted if in your proposal you can support the participants' qualifications to speak about the topic. Please submit your proposal online using this page (https://symposium2020.icsa.org/session-submission/) **before November 15, 2019**. The acceptance of invited sessions will be determined by **December 15, 2019**. In order to secure the invited session slot, the presenters will be required to register to the symposium and submit the abstracts online by **February 15, 2020**.

Call for Student Paper Award Applications

Up to eight student award winners (five Student Travel Awards, one Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper, and possible two ASA Biopharmaceutical Awards) will be selected. Each winner will receive a plaque or certificate, an award for travel and registration reimbursement up to $1,000 or a cash award of $550, whichever is bigger, as well as free registration for a short course. The deadline for applications is February 15, 2020.

ICSA 2020 China Conference (June 26 – 29, 2020)

The 2020 ICSA China Conference will be held at Zhongnan University of Economics and Law, Wuhan, China from June 26 to June 29, 2020. For information, please contact Scientific Program Committee Chair Professor Ying Zhang at ying.zhang@unmc.edu or Co-Chair Professor Hui Zhao at hzhao@zuel.edu.cn. Please visit http://ts.occamedu.com/?from=singlemessage&isappinstalled=0# for details including the key dates.

Call for Invited Session Proposals

The conference scientific program committee welcomes invited session proposals. An invited session consists of either 4 presenters or 3 presenters and 1 discussant. The one-talk rule will be applied (i.e., each speaker can only give one invited talk). It is required to confirm all speakers' availability before the proposal submission. Please send your proposal to one of the members of the Scientific Program Committee **before December 15, 2019**. The acceptance of invited sessions will be determined by **January 15, 2020**. In order to secure the invited session slot, the presenters will be required to register to the conference and submit the abstracts online by **March 15, 2020**.

Call for ICSA China Conference Junior Researcher Award Applications

The 2020 ICSA China Conference invites applications for **ICSA China Conference Junior Researcher Award.** Awardees will be selected from students or junior researchers who submit their papers for presentations at the 2020 ICSA China Conference and received their doctoral degrees no earlier than **March 1, 2014.** All qualified speakers are encouraged to submit a research paper in topics on either methodological research or novel application of statistical methods to real-world problems. Jointly authored papers are acceptable, but the applicant is expected to be the lead author and present the work in the meeting. The winners will be selected by the 2020 ICSA China Conference Junior Award Committee and the awards will be presented to the winners during the Banquet on June 27, 2020. To apply, please send an email to the Committee Chair **Dr. Xingqiu Zhao** at xingqiu.zhao@polyu.edu.hk by **March 1, 2020** with the subject title "Application - 2020 ICSA China Conference Junior Researcher Award" with the attached curriculum vita and paper of the completed research to be presented (both in pdf).

Scientific Program

December 20 8:30-9:00am

Openning Ceremony
Room: Crystal Ballroom

Chair:Hongzhe Li, University of Pennsylvania

December 20 9:00-10:00

Peter Hall Lecture
Room: Crystal Ballroom
Chair：T. Tony Cai, University of Pennsylvania
Organizer： ICSA 2019 Organizing Committee

9:00 Communication-Efficient Accurate Statistical Estimation
Jianqing Fan, Princeton University

December 20 10:30-12:10

S150: Causal inference and related methodology in health sciences
Room: R101
Chair: Guoshuai Cai, University of South Carolina
Organizer: Jian Wang, The University of Texas MD Anderson Cancer Center

10:30 A Bayesian semiparametric latent variable approach to causal mediation
Yisheng Li, The University of Texas MD Anderson Cancer Center

11:05 Challenge and promise of observational studies in cancer research
Yu Shen, The University of Texas MD Anderson Cancer Center

11:40 Bidirectional mediation to quantify direct and indirect effects with application to obesity and diabetes
Rajesh Talluri, University of Mississippi Medical Center

S029: Statistical machine learning in data science
Room: R102
Chair: Jingyi Jessica Li, University of California, Los Angeles
Organizer: Qing Zhou, UCLA Department of Statistics

10:30 Accurate and Efficient Machine Learning Methods
Ping Li, Baidu Research USA

10:55 Penalty Method for Variance Component Selection
Hua Zhou, UCLA

11:20 Identifiability of nonparametric mixture models, clustering, and semi-supervised learning
Nikhyl Aragam, University of Chicago

11:45 Directed acyclic graphs on network data
Qing Zhou, UCLA Department of Statistics

S034: False discovery rate methodology
Room:R103
Chair: Wenguang Sun, University of Southern California
Organizer:Ruth Heller, Tel-Aviv University

10:30 A Structure-Adaptive Learning Algorithm for Online False Discovery Rate Control
Wenguang Sun, University of Southern California

10:55 Controlling FDR while highlighting selected discoveries
Marina Bogomolov, Technion - Israel Institute of Technology

11:20 Simultaneous confidence intervals in sequential estimation
Aaditya Ramdas, Carnegie Mellon University

11:45 Practical aspects of using False Discovery Rate
Yoav Benjamini, Tel Aviv University

S006: Frontiers in Financial Statistics and Beyond
Room: R104
Chair: Yingying Li, Hong Kong University of Science and Technology
Organizer: Yingying Li, Hong Kong University of Science and Technology

10:30 Max-linear regression models with regularization
Zhengjun Zhang, University of Wisconsin

10:55 Factor Modeling for Volatility
Yi Ding, Hong Kong University of Science and Technology

11:20 Estimating Large Efficient Portfolios with Heteroscedastic Returns
Mengmeng Ao, Xiamen University

11:45 Specification Tests for Covariance Structures in High-Dimensional Statistical Models
Cheng Yong Tang, Temple University

S161: Recent Advances in Statistical Learning for Healthcare and Biomedical Problems
Room: R105
Chair: Ji Zhu, University of Michigan
Organizer: Ji Zhu, University of Michigan

10:30 Minorization-Maximization-based Boosting for Large-scale Survival Analysis with Time-Varying Effects
Zhi (Kevin) He, University of Michigan

10:05 Functional Regression for Brain Imaging
Bin Nan, University of California, Irvine

11:40 A Bayesian Approach to Joint Estimation of Multiple Graphical Models
George Michailidis, U of Florida

S036: Advanced Topics in Survival Analysis
Room: R108
Chair: Yair Goldberg, Technion
Organizer: Ruth Heller, Tel-Aviv University

10:30 Left without being seen: The disappearance of impatient patients, combining current-status, right-censored and left-censored data
Yair Goldberg, Technion - Israel Institute of Technology

10:55 Ventilation Prediction for ICU Patients with LSTM-based Deep Relative Risk Model
Bin Liu, Southwestern University of Finance and Economics

11:20 Analysis of semi-competing risks data via bivariate longitudinal models
Daniel Nevo, Tel Aviv University

11:45 Marginalized frailty-based illness-death model with application to biobank data
Malka Gorfine, Tel Aviv University

S041: Novel Statistical Approaches to Investigate Cancer Immunotherapy

Room: R109
Chair: Dongjun Chung, Medical University of South Carolina
Organizer: Dongjun Chung, Medical University of South Carolina

10:30 A Transcriptome Based Nonparametric Method to Deconvolute Immune Cells and Cancer Subtypes
Guoshuai Cai, University of South Carolina

10:55 A statistical framework to investigate molecular mechanisms associated with tumor microenvironment
Dongjun Chung, Medical University of South Carolina

11:20 Phase I/II dose finding interval design for Immunotherapy
Yeonhee Park, Medical University of South Carolina

11:45 Sparse LDA with Network-Guided Block Covariance Matrix
Jin Hyun Nam, Medical University of South Carolina

S142: Promoting Statistical Consulting and Collaboration in China

Room: R111
Chair: Xiaoyue Niu, Pennsylvania State University
Organizer: Ke Deng, Tsinghua University

10:30 Effective Communication for Successful Collaboration
Xiaoyue Niu, Pennsylvania State University

10:55 Statistical Consulting in the Era of Data Science
Lillian Lin, Retired Statistical Consultant

11:20 Strategies for promoting the engagement of students in statistical consulting
Ximing Xu, Nankai University

11:45 Statistical Consulting Practice at Tsinghua University
Mengzhao Gao, Tsinghua University

S082: Complex Innovative Designs in Practice of Early Phase Drug Development

Room: R201
Chair: Yuan Ji, The University of Chicago
Organizer: Sue-Jane Wang, Office of Translational Sciences

10:30 Model-based Phase I Designs with Incorporation of Individualized Dosing Using Toxicity Scores from Multiple Treatment Cycles
Jun Yin, Mayo Clinic

10:55 Innovative trials for early oncology drug development in China: opportunities& challenges
Xiang Guo, BeiGene

11:20 Rolling dose-finding designs
Daniel Li, Juno Therapeutics

11:45 A New Bayesian Framework for Master Protocols with Type I Error Control
Yuan Ji, The University of Chicago

S157: Deep learning and applications

Room: R202
Chair: Yuying Xie, Michigan State University
Organizer: Yuying Xie, Michigan State University
Co-Organizer: Matthew Hirn, Michigan State University

10:30 Invariant Data Representations with Multiscale Mathematical Models for ConvNets
Matthew Hirn, Michigan State University

10:55 Robustness and Sparsity in Deep Learning
Yuan Yao, Hong Kong University of Science and Technology

11:20 PDE-based Methods for Interpolation on High Dimensional Point Cloud
Zuoqiang Shi, Tsinghua University

11:45 Predicting Plant Stress Responses Using Deep Neural Network
Yuying Xie, Michigan State University

S059: Estimation from imperfectly observed data

Room: R301
Chair: Ronghui Xu, University of California
Organizer: Aurore Delaigle, University of Melbourne

10:30 Learning from EMR/EHR data to estimate treatment effects using high dimensional claims codes
Ronghui Xu, University of California

10:55 Variable selection and estimation in generalized linear models with measurement error
Liqun Wang, University of Manitoba

11:20 Estimating the covariance of fragmented functional data
Wei Huang, University of Melbourne

11:45 Density Estimation of Usual Intake for Food Consumption Data
Zhendong Huang, School of Mathematics and Statistics, The University of Melbourne

S058: Limit Theorems of Random Fields and Related Topics

Room: R302
Chair: Zhonggen Su, Zhejiang University
Organizer: Wensheng Wang, Hangzhou Dianzi University

10:30 Some recent results on multivariate Gaussian random fields
Yimin Xiao, Michigan State Univeristy

10:55 Derivatives of local times for some Gaussian fields
Fangjun Xu, East China Normal University

11:20 Probabilities of deviations for record numbers in random walks
Yuqiang Li, School of statistics, East China Normal University

11:45 The moduli of non-differentiability for Gaussian random fields with stationary increments
Wensheng Wang, Hangzhou Dianzi University

S147: Statistical Methods and Algorithms for High-dimensional Biomedical Data

Room: R303
Chair: Fei Zou, UNC-CH
Organizer: Fei Zou, UNC-CH

10:30 A Fast and powerful eQTL weighted method to detect genes associated with complex trait using GWAS summary data
Xuexia Wang, University of North Texas

11:05 Comparisons and Validation of the Three Pulmonary Nodule Malignancy Risk Models (Brock, Radiomics, Deep Learning): A Secondary Analysis of Data from the National Lung Screening Trial
Fenghai Duan, Brown University

11:40 Double Deep Learning for Adjusting Complex Confounding Structures In Observational Data
Fei Zou, UNC-CH

S151: The Use of Spectral Methods in Statistics: Theory and Applications

Room: R304
Chair: Yuting Wei, Carnegie Mellon University
Organizer: Tracy Ke, Harvard University

10:30 The Phase II data for the network of statisticians
Jiashun Jin, Carnegie Mellon University

10:55 Asymptotics of empirical eigenstructure for high dimensional spiked covariance
Weichen Wang, Two Sigma Investments

11:20 Detecting Rare and Weak Spikes in Large Covariance Matrices
Tracy Ke, Harvard University

11:45 A geometric perspective of hypothesis testing
Yuting Wei, Carnegie Mellon University

S064: New methods of testing and classification in complex data

Room: R305
Chair: Wanjie Wang, National University of Singapore
Organizer: Wanjie Wang, National University of Singapore

10:30 Innovated power enhancement for testing multi-factor pricing models with a large number of assets
Xiufan Yu, Pennsylvania State University

10:55 Topics on multiple testing
Hongyuan Cao, Florida State University

11:20 Hierarchical Community Detection with Fiedler Vectors
Xiaodong Li, UC Davis

11:45 Principal Boundary on Riemannian Manifolds and Classification Problem
Zhigang Yao, National University of Singapore

S072: Measuring and testing nonlinear dependence

Room: R306
Chair: Liping Zhu, Renmin University of China
Organizer: Liping Zhu, Renmin University of China

10:30 Test for conditional independence with application to conditional screening
Yeqing Zhou, Tongji University

10:55 A New Framework for Distance and Kernel-based Metrics in High Dimensions
Xianyang Zhang, Texas A&M University

11:20 Ball Covariance
Xueqin Wang, Sun Yat-sen University

11:45 Testing the Linear Mean and Constant Variance Conditions in Sufficient Dimension Reduction
Tingyou Zhou, Zhejiang University of Finance & Economics

S109: Statistical and Machine Learning Methods with Application in AI Transportation

Room: R307
Chair: Chengchun Shi, NC State University
Organizer: Rui Song, North Carolina State University

10:30 Statistics, Optimization and Deep Learning in the ride-sharing Industry
Fan Zhou, School of Statistics and Management, Shanghai University of Finance and Economics

11:05 A statistical and machine learning framework for new energy vehicle ride sharing system
Kaixian Yu, Didi Chuxing

11:40 Recent advances in landmark-based scalable spectral clustering
Guangliang Chen, San Jose State University

December 20 13:30-15:10

S099: Recent Advance in Bayesian Data Science

Room: R101
Chair: Haiying Wang, University of Connecticut
Organizer: Ming-Hui Chen, University of Connecticut

13:30 Bayesian Variable Selection with Application to High Dimensional EEG Data
Dipak Dey, Faculty

13:55 Registration Enabling Seamless Phase 1/2 Oncology Trial Design
Lei Cao, Changchun University of Technology

14:20 Bayesian Hierarchical Spatial Regression Models for Spatial Data in the Presence of Missing Covariates with Applications
Zhihua Ma, Shenzhen University

14:45 Bayesian Meta-Regression Hierarchical Models for Cholesterol Data
Ming-Hui Chen, University of Connecticut

S051: New Methodology for the Analysis of Neuroimaging Data

Room: R102
Chair: Todd Ogden, Columbia University
Organizer: Todd Ogden, Columbia University

13:30 A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome]{A Bayesian Approach to Joint Modeling of Matrix-valued Imaging Data and Treatment Outcome with Applications to Depression Studies
Bei Jiang, University of Alberta

13:55 Improved prediction of brain age using multimodal neuroimaging data
Fengqing (Zoe) Zhang, Drexel University

14:20 Multivariate Spline Estimation and Inference for Image-on-scalar Regression
Shan Yu, Iowa State University

14:45 Covariate Assisted Principal Regression for Covariance Matrix Outcomes with an Application to fMRI
Xi Luo, University of Texas Health Science Center at Houston

S021: Recent Statistical Advances in Biomedical Research

Room: R103
Chair: Lianfen Qian, Florida Atlantic University
Organizer: Lianfen Qian, Florida Atlantic University

13:30 Asymptotic distribution of the bias corrected LSEs in measurement error linear regression models under long memory
Hira Koul, Michigan State University

13:55 A Copula Model Approach for Regression Analysis of Informatively Interval-censored Failure Time Data
Jianguo Sun, University of Missouri

14:20 Personalized Treatment Selection for Joint Optimization of Survival and Other Outcomes
Somnath Datta, University of Florida

14:45 Oracally Efficient Estimation and Simultaneous Inference in Partially Linear Single-index Models for Longitudinal Data
Suojin Wang, Texas A&M University

S035: Multiple comparisons theory and applications

Room: R104
Chair: Fei Xue, University of Pennsylvania
Organizer: Ruth Heller, Tel-Aviv University

13:30 Closed testing and admissibility of procedures controlling false discovery proportions
Jelle Goeman, Leiden University Medical Center

14:05 All-Resolutions Inference for Brain Imaging
Aldo Solari, University of Milano-Bicocca

14:40 Extrapolating expected accuracy for large multi-class problems
Yuval Benjamini, Hebrew University of Jerusalem ISRAEL

S134: Recent Advances in Machine Learning
Room: R105
Chair: Ming Yuan, Columbia University
Organizer: Ming Yuan, Columbia University

13:30 Low-Tubal-Rank Tensor Recovery from One-bit Measurement
Jianjun Wang, Southwest University

13:55 The power of depth for deep nets in learning theory
Shao-bo Lin, Xi'an Jiaotong University

14:20 Estimation of the Number of Endmembers via Thresh olding Ridge Ratio Criterion
Xuehu Zhu, Xi'an Jiaotong University

14:45 A Fast and Accurate Frequent Directions Algorithm for Low Rank Approximation via Block Krylov Iteration
Yao Wang, Xi'an Jiaotong University

S077: Robust and efficient modern regression in high dimensional and complex data
Room: R106
Chair: Xiaodong Li, UC Davis
Organizer: Haochang Shou, University of Pennsylvania
Co-Organizer: Wanjie Wang, National University of Singapore

13:30 Total Variation Regularized Frechet Regression for Metric-Space Valued Data
Zhenhua Lin, National University of Singapore

13:55 Bayesian Covariate-dependent Gaussian Graphical Model
Yingying Wei, The Chinese University of Hong Kong

14:20 Pairwise-rank-likelihood methods for the semiparametric transformation model
Tao Yu, National University of Singapore

14:45 A constructive approach to L_0 penalized regression
Yanyan Liu, Wuhan University

S017: Innovative Statistical Methods for Analysis of EHR Data
Room: R108
Chair: Yize Zhao, Yale University
Organizer: Qi Long, University of Pennsylvania

13:30 Electronic Health Record Phenotyping using Anchor-Positive and Unlabeled Patients
Jinbo Chen, University of Pennsylvania

13:55 Data-driven discovery of medical terms from Chinese electronic health records
Sheng Yu, Tsinghua University

14:20 Functional clustering methods for extracting features from EHR biomarker history data
Jason Roy, Rutgers School of Public Health

14:45 Combining inverse-probability weighting and multiple imputation to adjust for selection bias due to missing data in EHR-based research
Sebastien Haneuse, Harvard T.H. Chan School of Public Health

S084: Statistical Methods for Large-Scale Networks
Room: R109
Chair: Junwei Lu, Harvard University
Organizer: Junwei Lu, Harvard University

13:30 Nonregular and Minimax Estimation of Individualized Thresholds in High dimension with Binary Responses
Yang Ning, Cornell University

13:55 Machine Learning Methods For Estimation and Inference in Differential Networks
Mladen Kolar, University of Chicago

14:20 Network Clustering Hypothesis Testing
Junwei Lu, Harvard University

14:45 Estimating Joint Latent Space Models for Network Data with High-Dimensional Node Variables
Xuefei Zhang, University of Michigan

S135: Random Matrix Theory and its Applications to Statistics
Room: R110
Chair: Fang Han, University of Washington
Organizer: Fang Han, University of Washington

13:30 Eigenvector distribution of deformed random matrices
Xiucai Ding, Duke University

13:55 ON EIGENVALUES OF A HIGH-DIMENSIONAL SPATIAL-SIGN COVARIANCE MATRIX
Weiming Li, Shanghai University of Finance and Economics

14:20 On testing high dimensional white noise
Zeng Li, Southern University of Science and Technology

14:45 Can we trust PCA on nonstationary data?
Yanrong Yang, Australian National University

S003: Methodologies for complex survival data
Room：R111
Chair：Chin-Tsang Chiang, National Taiwan University
Organizer： Ming-Yueh Huang, Institute of Statistical Science, Academia Sinica

13:30 Information Synthesis and Variable Selection Using A Penalized Empirical Likelihood Approach
Ying Sheng, University of California at San Francisco

13:55 Synthesizing Independent Stagewise Trials for Optimal Dynamic Treatment Regimes
Yuan Chen, Columbia University

14:20 Quantile Residual Life Regression Based on Semi-Competing Risks Data
Jin-Jian Hsieh，National University of Singapore

14:45 Dimension reduction in multivariate baseline proportional hazards models
Ming-Yueh Huang, Institute of Statistical Science, Academia Sinica

S046: Recent method and technique developments in genomics and drug safety
Room: R201
Chair: Ming Wang, Pennsylvania State University
Organizer: Ming Wang, Pennsylvania State University

13:30 Group-level network inference via l_0 shrinkage and graph

combinatorics
Shuo Chen, University of Maryland, School of Medicine

13:55 Integrated sequencing analysis for virus detection in Human disease
Lijun Zhang, Pennsylvania State University College of Medicine

14:20 Statistical test of structured continuous trees based on discordance matrix
Lin Wan, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

14:45 Bayesian Modeling of Rare Events Data with Missing Not At Random
Shouhao Zhou, The Pennsylvania State University

S115: Recent developments in discriminant and multivariate analysis

Room: R202
Chair: Binyan Jiang, The Hong Kong Polytechnic University
Organizer: Binyan Jiang, The Hong Kong Polytechnic University
Co-Organizer: Chenlei Leng, University of Warwick

13:30 A unified frameowork for sufficient dimension reduction in high dimensions
Qing Mai, Florida State University

14:05 Robust Principal Component Analysis by Manifold Optimization
Teng Zhang, University of Central Florida

14:40 Linear discriminant analysis with high dimensional mixed variables
Binyan Jiang, The Hong Kong Polytechnic University

S087: Innovative study designs and analyses for early-phase clinical trials

Room: R301
Chair: Yunda Huang, Fred Hutch
Organizer: Xiwu Lin, Johnson & Johnson
Co-Organizer:Yunda Huang, Co-Organizer Affiliation

13:30 The paradox of increasing sample size and decreasing power in testing the difference between two independent binomial proportions
Youyi Fong, University of Washington, Department of Biostatistics

13:55 Registration Enabling Seamless Phase 1/2 Oncology Trial Design
Jun Dong, Amgen Biopharmaceutical R&D (Shanghai) Co., Ltd.

14:20 Review and Examples of Master Protocols
Li Li, R&G Pharma

14:45 Discussant
Yunda Huang, Fred Hutchinson Cancer Research Center

S162: Statistical models for diseases with spatial or temporal variations

Room: R302
Chair: Huiyan Sang, Texas A&M University
Organizer: Le Bao, Pennsylvania State Unmioversity
Co-Organizer: Richard Li, Yale University

13:30 Modeling heroin-related EMS calls in space and time
Zehang Li, Yale School of Public Health

14:05 An ensemble approach to predicting the impact of vaccination on rotavirus disease in Niger
Jaewoo Park, Yonsei University

14:40 Kernel Machine and Distributed Lag Models for Assessing Windows of Susceptibility to Mixtures of Time-Varying Environmental Exposures in Children's Health Studies
Ander Wilson, Colorado State University

S148: Statistical Issues in Imaging Data Analysis

Room: R303
Chair: Bin Nan, University of California, Irvine
Organizer: Bin Nan, University of California, Irvine

13:30 Bayesian Spatial Blind Source Separation via Thresholded Gaussian Processes
Jian Kang, University of Michigan

14:05 High-Dimensional Spatial Quantile Function-on-Scalar Regression in Neuroimaging Analysis
Linglong Kong, University of Alberta

14:40 A Bayesian State-Space Approach to Mapping Directional Brain Networks
Tingting Zhang, The University of Virginia

S121: Complex data analysis and its applications

Room: R304
Chair: Sai Li, University of Pennsylvania
Organizer: Zijian Guo, Rutgers University

13:30 Joint modeling of multivariate continuous and time-to-event data
Xinyuan Song, Chinese University of Hong Kong

13:55 Statistical learning for individualized asset allocation
Yingying Li, Hong Kong University of Science and Technology

14:20 Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach
Sai Li, University of Pennsylvania

14:45 Generalized integration model for Improved Statistical Inference by Leveraging External Summary Data
Kai Yu, National Cancer Institute

S030: Statistical Inference for High-dimensional Tensor Data

Room: R305
Chair: Zijian Guo, Rutgers University
Organizer: Anru Zhang, University of Wisconsin-Madison

13:30 Tensor bandits for online interactive recommendation
Will Wei Sun, Purdue University

14:05 Subspace-regularized Tensorial Parameter Estimation
Xin Zhang, Florida State University

14:40 High-dimensional Tensor Regression Analysis
Anru Zhang, University of Wisconsin-Madison

S037: New Analytical Solutions for Single-Cell and Functional Genomic Data

Room: R306
Chair: Hongkai Ji, Johns Hopkins Bloomberg School of Public Health
Organizer: Hongkai Ji, Johns Hopkins Bloomberg School of Public Health

13:30 A statistical simulator scDesign for rational scRNA-seq experimental design
Jingyi Jessica Li, UCLA

14:05 Single-Cell Transcriptome and Regulome Data Integration
Weiqiang Zhou, Johns Hopkins Bloomberg School of Public Health

14:40 SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing
Yuchao Jiang, University of North Carolina at Chapel Hill

S057: Statistical Analysis of Complex Data
Room: R307
Chair: Xingqiu Zhao, The Hong Kong Polytechnic University
Organizer: Jianguo Sun, University of Missouri

13:30 Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression
Hui Zhao, Zhongnan University of Economics and Law

13:55 A Vine Copula Approach for Regression Analysis of Bivariate Current Status Data with Informative Censoring
Huiqiong Li, Yunnan University

14:20 Personalized Glucose Prediction Using Attention-based RNN
Ran Duan, Eli Lilly and Company

14:45 Regression analysis of informatively interval-censored failure time data with semiparametric linear transform ation model
Da Xu, Center for Applied Statistical Research, School of Mathematics, Jilin University

December 20 15:40-17:20

S002: Recent Advances in Functional Data Analysis
Room: R101
Chair: Jeng-Min Chiou, Academia Sinica
Organizer: Jeng-Min Chiou, Academia Sinica

15:40 Additive Regression for Predictors of Various Natures and Hilbertian Responses with Application to Censored and Missing Data
Byeong Park, Seoul National University

16:05 Predictive Functional Linear Models with Semiparametric Single-Index Interactions
Naisyin Wang, University of Michigan

16:30 A new knn-classifier for functional data with applications
Jin-Ting Zhang, National University of Singapore

16:55 Hypothesis Testing in Large-scale Functional Linear Regression
Kaijie Xue, Nankai university

S010: New challenges in nonparametric inference
Room: R102
Chair: Young Kyung Lee, Kangwon National University
Organizer: Young Kyung Lee, Kangwon National University

15:40 Nonparametric modeling of heteroscedasticity in multi-dimensional regression
Kyusang Yu, Konkuk University

16:05 Tail estimation for the spectral density matrix of multivariate Gaussian random fields
Chae Young Lim, Seoul National University

16:30 Inference of break-points in high-dimensional time series
Weining Wang, City University of London

16:55 Bias Reduction for Nonparametric and Semiparametric Regression Models
Ming-Yen Cheng, Hong Kong Baptist University

S005: Machine Learning Methods in Biomedical Science
Room: R103
Chair: Fenghai Duan, Brown University
Organizer: Rui Feng, University of Pennsylvania

15:40 Data Integration of Multiple Genome-Wide Association Studies Under Group Homogeneous Structure
Yuan Jiang, Oregon State University

16:05 A unified machine learning method of determining the minimal important difference with the linear structure
Jiwei Zhao, State University of New York at Buffalo

16:30 A Bayesian Semi-supervised Approach to Key Phrase Extraction with Only Positive and Unlabeled Data
Sherry Wang, Southern Methodist University

16:55 Deep Learning with Graph Structure in Small Samples
Rui Feng, University of Pennsylvania

S088: Statistical Methods for Genomic and Transcriptomic Data Analysis
Room: R104
Chair: Wei Sun, Fred Hutch
Organizer: Wei Sun, Fred Hutch

15:40 Detecting Dense and Sparse Signals in Omics Studies
Chi Song, The Ohio State University

16:05 Dynamic Correlation Analysis for Omics Data
Tianwei Yu, Emory University

16:30 Statistical model for background mutation rate in cancer genomes
Lin Hou, Tsinghua University

16:55 scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition
Ruibin Xi, Peking University

S016: Inference with Complex Data
Room: R105
Chair: Qiwei Yao, London School of Economics
Organizer: Qiwei Yao, London School of Economics

15:40 Jump or Kink: Super-efficiency in Segmented Linear Regression Break-point Estimation
Yining Chen, London School of Economics

16:05 Model-based Outlier Detection in Multivariate Data with Applications to Detecting Cheating in Tests
Yunxiao Chen, London School of Economics and Political Science

16:30 A computationally efficient approach to the multivariate changepoint problem
Idris Eckley, Lancaster University

16:55 Online High Dimensional Covariance Change Point Detection
Clifford Lam, London School of Economics and Political Science

S066: Recent Advances in Statistical Learning
Room: R106
Chair: Yin Xia, Fudan University
Organizer: Yin Xia, Fudan University

15:40 Adaptive Design of Network A/B tests
Feifang Hu, George Washington University

16:05 Support vector machine in construction of personal treatment rules
Peter Song, University of Michigan

16:30 Random projection pursuit regression for high-dimensional complex data
Sijian Wang, Rutgers University

16:55 Ultrahigh Dimensional Precision Matrix Estimation via Refitted Cross Validation
Zhao Chen, Fudan University

S160: Causal Inference

Room: R107
Chair: Zhi Geng, Peking University
Organizer: Jinzhu Jia, Peking University

15:40 Estimation of Optimal Individualized Treatment Rule Using the Covariate-Specific Treatment Effect Curve with High-dimensional Covariates
Xiaohua Zhou, Beijing International Center for Mathematical Research and Department of Biostatistics, Peking University

16:05 Specification tests for generalized propensity scores using double projections
Xiaojun Song, Peking University

16:30 Covariate Adjustment in Completely Randomized Experiments With Noncompliance
Hanzhong Liu, Tsinghua University

16:55 On the Efficiency of Logistic Regression Estimators in Estimating The Causal Effect
Jinzhu Jia, Peking University

S042: Recent Advancement in Biostatistics Methodology

Room: R108
Chair: Chen Hu, Johns Hopkins University
Organizer: Mei-Cheng Wang, Department of Biostatistics, Johns Hopkins University

15:40 Joint analysis of multiple longitudinal and survival data measured on nested time-scales: an application to predicting infertility
Rajeshwari Sundaram, Speaker Affiliation

16:05 Modeling and Correlation Estimation for Bivariate Recurrent Event Processes
Mei-Cheng Wang, Department of Biostatistics, Johns Hopkins University

16:30 Analysis of competing risks data with dependent truncation
Yu-Jen Cheng, National Tsing Hua University

16:55 Biomarker Guided Phase II Two-Stage Design for Targeted Therapy
Zheyu Wang, JHU

S168: Complex Data Analysis in Business, Economics and Industry

Room: R109
Chair: Ching-Kang Ing, National Tsing Hua University
Organizer: Ching-Kang Ing, National Tsing Hua University

15:40 Quality Big Data
Fugee Tsung, HKUST

16:05 On selecting valid instruments for structural vector autoregression
Chor-Yiu Sin, National Tsing Hua University

16:30 Mean squared prediction errors of integrated autoregressive models with polynomial time trends
Shu-Hui Yu, National University of Kaohsiung

16:55 Hing-dimensional model selection under covariate shift
Ching-Kang Ing, National Tsing Hua University

※Discussant
Jun Wang, CFDA

S155: Recent advances in biomedical big data analytics

Room: R110
Chair: Zhixiang Lin, The Chinese University of Hong Kong
Organizer: Zhixiang Lin, The Chinese University of Hong Kong

15:40 MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting linkage disequilibrium and horizontal pleiotropy
Jin Liu, Duke-NUS Medical School

16:05 Model-based microbiome data ordination: A variational approximation approach
Tao Wang, Shanghai Jiao Tong University

16:30 Post-GWAS data integration identifies risk factors for Alzheimer's disease
Qiongshi Lu, University of Wisconsin-Madison

16:55 Statistical methods for data integration in single-cell genomics
Zhixiang Lin, The Chinese University of Hong Kong

S044: Data science and statistics in IT companies

Room: R111
Chair: Lu Zhang, google
Organizer: Lu Zhang, google

15:40 Semantic clustering of YouTube videos
Ying Liu, Google

16:05 Comparison Studies of Multi-Armed Bandit Algorithms for Display Advertising Optimization
Wanghuan Chu, Google

16:30 Inferring Impact Direction Graphs from Large Scale Online User Engagement Data
Yuxiang Xie, Snap Inc.

16:55 Statistics and Big data at Google
Lu Zhang, google

S100: New Advances on Statistical Modeling of Complex Data

Room: R201
Chair: Dipak Dey, University of Connecticut
Organizer: Victor Hugo Lachos Davila, University of Connecticut

15:40 Mixtures of factor analysis models with covariates for multiply censored dependent data
Tsung I-Lin, National Chung Hsing University

16:05 Analysis of Multivariate Longitudinal Data with Censored and Intermittent Missing Responses
Wan-Lun Wang, Speaker Affiliation

16:30 Bayesian Analysis of Survival Data with Missing Censoring Indicators
Mauricio Castro, Pontificia Universidad Catolica de Chile

16:55 Likelihood-based Inference for Mixed-Effects Models with Censored Response Using Skew-Normal Distribution
Victor Hugo Lachos Davila, University of Connecticut

S111: Strategic and Statistical Considerations in Early Phase Drug Development

Room: R202
Chair: Fan Xia, BeiGene
Organizer: Fan Xia, BeiGene

15:40 Proof of Concept Decision Making in Phase 1b Cohort Expansion Study
Fei Ji, Eli Lilly and Company

16:05 Consideration of platform design in oncology development in China
Naitee Ting, Boehringer-Ingelheim

16:30 Bayesian Basket trial Design Accounting for Multiple Cutoffs of the Ambiguous Biomarker
Jin Xu, East China Normal University

16:55 Recent Advances for the Design and Conduct of Efficient Phase I/II Trials
J.Jack Lee, MD Anderson Cancer Center

S144: Bayesian Statistics

Room: R301
Chair: Suman Guha, Department of Statistics, Presidency University
Organizer: Anil Ghosh, Indian Statistical Institute
Co-Organizer: Subhajit Dutta, IIT KANPUR

15:40 Efficient Bernoulli factory MCMC for intractable likelihoods
Dootika Vats, Indian Institute of Technology

16:05 Estimating densities with nonlinear support using Fisher-Gaussian kernels
Kiranmoy Das, Indian Institute of Technology Kanpur

16:30 Estimating densities with nonlinear support using Fisher-Gaussian kernels
Minerva Mukhopadhyay, Indian Institute of Technology Kanpur

16:55 Bayesian Analysis with Gaussian Random Functional Dynamic Spatio-Temporal Model
Suman Guha, Department of Statistics, Presidency University

S127: Advances in Large Scale Data Analysis

Room: R302
Chair: Hua Zhou, UCLA
Organizer: Gang Li, UCLA

15:40 Bayesian Analysis of Multidimensional Functional Data
Donatello Telesca, UCLA

16:05 A New Joint Screening Method for Right-Censored Time-to-Event Data with Ultra-high Dimensional Covariates
Yi Liu, Ocean University of China

16:30 Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression
Hui Zhao, Zhongnan University of Economics and Law

16:55 Dimension Reduction via cross-validation metric learning
Linlin Dai, Southwestern University of Finance and Economics

S122: Traditional statistical techniques in new data setting

Room: R303
Chair: Binyan Jiang, The Hong Kong Polytechnic University
Organizer: Catherine Liu, The Hong Kong Polytechnic University

15:40 Modeling Count Time Series via Common Factors
Fangfang Wang, Worcester Polytechnic Institute

16:05 Variable Selection for Multiple Types of High-Dimensional Features With Missing Data
Kin Yau Wong, The Hong Kong Polytechnic University

16:30 High-order Imaging Regression via Internal Variation
Long Feng, City University of Hong Kong

16:55 Distributed Learning with Minimum Error Entropy Principle
Xin Guo, The Hong Kong Polytechnic University

S159: Recnet Develpments in Statistical Network Analysis

Room: R304
Chair: Ji Zhu, University of Michigan
Organizer: Ji Zhu, University of Michigan

15:40 Hierarchical community detection by recursive partitioning
Tianxi Li, University of Virginia

16:05 Popularity-Adjusted Block Models for Networks with Community Structure
Yuguo Chen, University of Illinois at Urbana-Champaign

16:30 Network Differential Connectivity Analysis
Ali Shojaie, University of Washington

16:55 Edgeworth approximation to network U-statistics
Yuan Zhang, Ohio State University

S075: Massive regression analysis

Room: R305
Chair: Binhuan Wang, NYU School of Medicine
Organizer: Huazhen Lin, Southwestern University of Finance and Economics

15:40 Additive partially linear models for massive heterogeneous data
Binhuan Wang, NYU School of Medicine

16:05 Discrepancy between global and local principal component analysis on large-panel high-frequency data
Xinbing Kong, Nanjing Audit University

16:30 Spatial-Temporal Prediction of PM2.5 concentration in North China Plain Using the Machine Learning
Bin Guo, Southwestern University of Finance and Economics

16:55 Identifying sensitive subset based on ultra-high dimensional correlated covariates for survival data
Ye He, University of Electronic Science and Technology of China

S103: Selective Inference and Multiple Comparisons

Room: R306
Chair: Hidetoshi Shimodaira, Kyoto University / RIKEN AIP
Organizer: Hidetoshi Shimodaira, Kyoto University / RIKEN AIP

15:40 Selective Inference after Unsupervised Hidden-Structure Identification
Ichiro Takeuchi, Nagoya Institute of Technology / RIKEN AIP

16:05 Perturbation of the expected Minkowski functional and its applications
Satoshi Kuriki, The Institute of Statistical Mathematics

16:30 Adjusting the bias of multiple testing by bootstrap resampling with “negative” sample size and its applications to phylogenetics
Hidetoshi Shimodaira, Kyoto University / RIKEN AIP

16:55 Selective inference for the problem of regions via multiscale bootstrap with applications to clustering and regression
Yoshikazu Terada, Osaka University

S139: Time Series Analysis

Room: R307
Chair: Shifeng Huang, Chair Affiliation
Organizer: Guangming Pan, Nanyang Technological University

15:40 Modeling Financial Time Series with Soft Information
Shih-Feng Huang, National University of Kaohsiung

16:05 Time Series Analysis with Unsupervised Learning
Meihui Guo, Dept. of Applied Mathematics, National Sun Yat-sen University

16:30 Symbolic Interval-Valued Data Analysis for Time Series based on Normality Assumption
Liang-Ching Lin, National Cheng Kung University

16:55 Clt for largest eigenvalues in high-dimensional nonstationary time series and its applications
Bo Zhang, University of Science and Technology of China

December 21 9:00-10:00

Pao-Lu Hsu Award Lecture and Award Ceremony

Room: Crystal Ballroom
Chair：Heping Zhang, Yale University
Organizer： ICSA 2019 Organizing Committee

9:00 Award ceremony: New Researcher Award and Pao-Lu Hsu Award

9:10 Fisher's 1918 Quantitative Genetics Model In the Genomics Era
Hongyu Zhao, Yale School of Public Health

December 21 10:30-12:10

S119: Recent Advances in High dimensional Statistics

Room: R101
Chair: Linjun Zhang, Rutgers University
Organizer: Zijian Guo, Rutgers University

10:30 Limit distribution theory in multiple isotonic regression
Cun-Hui Zhang, Rutgers University

10:55 Bayesian variance estimation in the Gaussian sequence model with partial information on the means
Gianluca Finocchio, University of Twente

11:20 Large Covariance Regression for Spatial Data
Wei Lin, Peking University

11:45 The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy
Linjun Zhang, Rutgers University

S055: Random Matrices Theory and Applications

Room: R102
Chair: Yuxin Chen, Princeton University
Organizer: Harrison Zhou, Yale University

10:30 Spectral graph matching and regularized quadratic relaxations
Zhou Fan, Yale University

10:55 Conformal prediction with localization
Leying Guan, Yale University

11:20 Optimality of Spectral Clustering for Gaussian Mixture Model
Anderson Zhang, The Wharton School, University of Pennsylvania

11:45 Adaptation in multivariate log-concave density estimation
Kyoung Hee Arlene Kim, Korea University

S076: Special Invited Session - Data Science Research at DiDi

Room: R103
Chair: Hongtu Zhu, Didi
Organizer: Hongtu Zhu, Didi and University of North Carolina

10:30 Challenges in Analyzing Two-sided Market and Its Application on Ride-sourcing Platform
Hongtu Zhu, Didi and University of North Carolina

11:05 Marry Data and AI with SQLFLow
Ziyao Gao (Jeremy), DiDi

11:40 Driving Risk Assessment for Ride-hailing Drivers
Liang Shi, Virginia Tech Transportation Institute

S007: Recent Advances on Large Complex Data

Room: R104
Chair: Xinghua Zheng, Hong Kong University of Science and Technology
Organizer: Yingying Li, Hong Kong University of Science and Technology

10:30 Information Based Complexity of High Dimensional Sparse Functions
Ming Yuan, Columbia University

10:55 Inference for Case Probability in High-dimensional Logistic Regression
Zijian Guo, Rutgers University

11:20 Statistical learning for individualized asset allocation
Rui Song, North Carolina State University

11:45 Change-detection-assisted multiple testing for spatiotemporal data
Lilun Du, Hong Kong University of Science and Technology

S065: Recent Advances in Statistical Theories and Applications

Room: R105
Chair: Yin Xia , Fudan University
Organizer: Yin Xia, Fudan University

10:30 Estimation for Double-Nonlinear Cointegration
Qiwei Yao, London School of Economics

10:55 Classification with imperfect training labels
Richard Samworth, University of Cambridge

11:20 Generative Link Prediction for Incomplete Networks with Node Features
Ji Zhu, University of Michigan

11:45 Fourier Transform Approach for Inverse Dimension Reduction Method
Xiangrong Yin, University of Kentucky

S136: Robust Statistics

Room: R108
Chair: Fang Han, University of Washington
Organizer: Fang Han, University of Washington

10:30 Confidence intervals for multiple isotonic regression and other monotone models
Qiyang Han, Rutgers University

10:55 Adaptive Minimax Density Estimation for Huber's Contamination Model under L_p losses
Zhao Ren, University of Pittsburgh

11:20 Consistency of a range of penalised cost approaches for detecting multiple changepoints
Chao Zheng, Lancaster University

11:45 On Perfect Classification and Clustering for Gaussian Processes
Subhajit Dutta, IIT KANPUR

S025: Recent Advances in Statistical Genomics

Room: R109
Chair: Hui Jiang, University of Michigan
Organizer: Hui Jiang, University of Michigan

10:30 A sparse clustering algorithm for identifying cluster changes across conditions with applications in single-cell RNA-sequencing data
Jun Li, University of Notre Dame

11:05 Statistical Analysis of Spatial Expression Pattern for Spatially Resolved Transcriptomic Studies
Xiang Zhou, University of Michigan

11:40 Dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder

Jin Gu, Tsinghua University

S033: Handling Complex Featured Data: Methods and Applications

Room: R110
Chair: Xin Liu, Shanghai University of Finance and Economics
Organizer: Grace Yi, University of Waterloo, University of Western Ontario

10:30 Support Vector Machine with Graphical Network Structures in Features
Wenqing He, University of Western Ontario

10:55 Analysis of multivariate longitudinal data from eyes - microperimetry macular sensitivity loss in patients with Stargardt Disease
Xiangrong Kong, Johns Hopkins University

11:20 Dynamic risk prediction of a clinical event with sparse and irregularly measured longitudinal biomarkers
Yayuan Zhu, University of Western Ontario

11:45 Degradation in common dynamic environments
Zhisheng Ye, National University of Singapore

S085: Leadership and Innovation in Drug Development Through Quantitative Research

Room: R111
Chair: Jianing Di, Janssen Research & Development
Organizer: Wei Shen, Eli Lilly and Company
Co-Organizer: Guohua (James) Pan, Janssen Research & Development

Discussants: *Xiaoni Liu*, Novartis
Haoda Fu, Eli Lilly and Company
Jianing Di, Janssen Research & Developmet
Ping Yan, Jiangsu Hengrui Medicine
Tony (Xiang) Guo, BeiGene

S079: Statistical Advancements for Emerging Challenges in Health Data Science

Room: R201
Chair: Wendy Lou, University of Toronto
Organizer: Wendy Lou, University of Toronto
Co-Organizer: Pingzhao Hu, University of Manitoba

10:30 Identifying the Best Predictive SNP in GWAS for Companion Diagnostics
Xinping Cui, Department of Statistics, University of California

10:55 Statistical approaches for identifying biomarkers for a group of cancer drugs
Jian Zhang, University of Kent

11:20 Computational methods to elucidate chromatin topological structure using 3D genomic maps
Shihua Zhang, Academy of Mathematics and Systems Biology，CAS

11:45 Deep learning for decoding molecular phenotypes with radiogenomics in breast cancer
Pingzhao Hu, University of Manitoba

S054: Statistical methodologies in clinical trials

Room: R202
Chair: Bingzhi Zhang, Sanofi
Organizer: Hui Quan, Sanofi

10:30 Efficient Sample Size Adaptation Strategy With Adjustment Of Randomization Ratio
Yijie Zhou, Vertex Pharmaceuticals

10:55 Optimal adaptive group sequential design with flexible timing of sample size determination
Bo Yang, Vertex

11:20 Historical data borrowing from multiple historical trials with commensurate prior
Bingzhi Zhang, Sanofi

11:45 Utilization of Robust Estimates of treatment Effect via Semi-Parametric Models in MRCT
Ming Tan, Georgetown University

S081: Design and Analysis for Complex Medical Studies

Room: R301
Chair: Haiyan Zheng, Newcastle University
Organizer: Yiyi Chen, Oregon Health & Science University

10:30 Robust Design Approaches in Biomedical Research
Timothy O'Brien, Loyola University Chicago

11:05 Analyzing longitudinal activity data collected from wearable devices
Meike Niederhausen, Oregon Health & Science University

11:20 Bayesian single-index joint models of multivariate longitudinal and survival data
An-Min Tang, Yunnan University

11:45 A case study of refining testing strategy using graphical approach
Eva Hua，Novartis (Shanghai)

S097: Recent Development on Missing Data Issues under Estimand Framework

Room: R302
Chair: Mouna Akacha, Chair Affiliation
Organizer: Jiawei Wei, Novartis

10:30 Estimands, Missing Data, and Sensitivity Analysis
Geert Molenberghs, I-BioStat, Hasselt University & KU Leuven

10:50 Cox Regression with Survival-Dependent Missing Covariate Values
Jun Shao, East China Normal University

11:10 The Challenges of Analyzing Drug Safety Data with Competing Risk Events and Some Thoughts
Aileen Zhu, China Novartis Institutes for BioMedical Research Co., Ltd.

11:30 Mixture of multivariate t linear Mixed Models With Missing Information
Tzy-Chy Lin, Center for Drug Evaluation

11:50 Discussant
Jun Wang, CFDA

S011: Modeling and analysis of spatial point pattern data

Room: R303
Chair: Chae Young Lim, Seoul National University
Organizer: Chae Young Lim, Seoul National University

10:30 Global multivariate point pattern models for rain type occurrence
Mikyoung Jun, Texas A&M University

11:05 Spatial Sampling Design using Generalized Neyman-Scott Process
Zhengyuan Zhu, Iowa State University

11:40 Intensity estimation for spatial point processes
Ottmar Cronie, Umeå University

S145: Special Invited Paper of Statistics and Its Inferenece
Room: R304
Chair: Minghui Chen, University of Connecticut
Organizer: Minghui Chen, University of Connecticut
Yuedong Wang, University of California, Santa Barbara

10:30 Doubly Regularized Estimation and Selection in Linear Mixed-Effects Models for High-Dimensional Longitudinal Data
Ling Zhou, Southwestern University of Finance and Economics

11:05 Optimal Treatment Assignment for Multiple Treatments with Analysis of Variance Decomposition
Zhilan Lou, Zhejing University of Finance and Economics

11:40 Bayesian Modeling and Uncertainty Quantification for Descriptive Social Networks
Sudipto Banerjee, University of California, Los Angeles

S140: High dimensional change point detection, Regression and Clustering
Room: R305
Chair: Guangming Pan, Nanyang Technological University
Organizer: Guangming Pan, Nanyang Technological University

10:30 A Composite Likelihood-based Approach for Change-point Detection in Spatio-temporal Process
Chun Yip Yau, Chinese University of Hong Kong

10:55 Distributed linear regression in high dimensions
Yue Sheng, University of Pennsylvania

11:20 High dimensional clustering
Guangming Pan, Nanyang Technological University

11:45 Selection of the number of change-points via error rate control
Changliang Zou, Nankai University

S001: Complex Medical Data Analysis
Room: R306
Chair: Hua Liang, George Washington University
Organizer: Hua Liang, George Washington University

10:30 Efficient estimation of the Nonparametric Mean and Covariance Functions for Longitudinal and Sparse Functional Data
Ling Zhou, Southwestern University of Finance and Economics

10:55 Bayesian Piecewise Linear Mixed Models with a Random Change Point
Xiang Liu, University of South Florida

11:20 Fiducial Model Selection
Xinmin Li, Qingdao University

11:45 High-dimensional Tobit models
Hua Liang, George Washington University

S167: New Advances on Complex Data Analysis
Room: R307
Chair: Pengsheng Ji, University of Georgia
Organizer: Pengsheng Ji, University of Georgia

10:30 Penalized Empirical Likelihood for the Sparse Cox Model
Yichuan Zhao, Georgia State University

10:55 Retrospective score tests versus prospective score tests for genetic association with case-control data
Yukun Liu, East China Normal University

11:20 Brain-wide organizations of neuronal activity in larval zebrafish
Yu Hu, The Hong Kong University of Science and Technology

11:45 Interaction Pursuit Biconvex Optimization
Yuehan Yang, Central University of Finance and Economics

December 21 13:30-15:10

S124: Novel approaches for analysis of probability and non-probability samples
Room: R101
Chair: Qixuan Chen, Columbia University
Organizer: Qixuan Chen, Columbia University

13:30 Combining Probability Non-probability Samples: Theory and Practice
Michael Elliott, University of Michigan

13:55 Bootstrap inference for the finite population total under complex sampling designs
Zhonglei Wang,Xiamen University

14:20 Methodologies for Analyzing Non-probability Survey Samples
Changbao Wu, University of Waterloo

14:45 Bayesian Inference for Sample Surveys in the Presence of High-Dimensional Auxiliary Information
Qixuan Chen, Columbia University

S028: Statistical and Computational Genomics
Room: R102
Chair: Qing Zhou, UCLA Department of Statistics
Organizer: Qing Zhou, UCLA Department of Statistics

13:30 Statistical methods for high-resolution chromosome conformation data analysis
Wenxiu Ma, University of California Riverside

13:55 Efficient algorithms for resampling-based hypothesis testing in genomic data analysis
Hui Jiang, University of Michigan

14:20 Personalized Risk Predictions with Deep Learning Methods in the Presence of Missing and Biased Electronic Health Record and Genomics Data
Hua Zhong, NYU Langone Health

14:45 A linear mixed model framework to study gene-environment interactions
Ku, Hung-Chih，DePaul University

S012: Analysis of Semiparametric Models
Room: R103
Chair: Qizhai Li, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Organizer: Liuquan Sun, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

13:30 Inference in a mixture additive hazards cure model
Haijin He, Shenzhen University

14:05 Semiparametric Inference for the Functional Cox Model
Meiling Hao, University of International Business and Economics

14:40 Efficient Fused Learning for Distributed Imbalanced Data*Jie Zhou*, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

S049: Utilization of Big Data in Precision Medicine
Room: R104
Chair: Zhengjia Chen, Emory University

Organizer: Zhengjia Chen, Emory University

13:30 Subgroup Discovery Using Consensus Clustering Methods
J. Richard Landis, University of Pennsylvania

13:55 Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis
Qi Long, University of Pennsylvania

14:20 An Unified Framework of Personalized Network Recovery and Detection
Ming Wang, Pennsylvania State University

14:45 Estimation of personalized maximum tolerated dose (pMTD) by incorporation of patient's genomic profiles and all toxicity information in cancer Phase I clinical trial
Zhengjia Chen, Emory University

S163: New development for statistical analysis

Room: R105
Chair: Qihua Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Organizer: Qihua Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

13:30 Oracally Efficient Estimation and Simultaneous Inference in Partially Linear Single-index Models for Longitudinal Data
Suojin Wang, Texas A&M University

14:05 Modeling and Analysis of Correlated Data using Pairwise Likelihood
Grace Yi, University of Waterloo, University of Western Ontario

14:40 On the asymptotic distribution of model averaging based on information criterion
Guohua Zou, Capital Normal University

S032: Recent Advances in Lifetime Data Analysis

Room：R107
Chair：Yong Chen, University of Pennsylvania
Organizer：Mei-Ling Ting Lee, University of Maryland

13:30 Semiparametric regression analysis for serial gap times with competing events
Shu-Hui Chang,National Taiwan University

13:55 High dimensional data reduction in risk and survival data analysis
Catherine Huber, University Paris Descartes

14:20 Statistical Analysis of Event Duration with Missing Origin
Yi Xiong, Simon Fraser University

14:45 Variable screening with multiple studies and its application in survival analysis
Tianzhou Ma, University of Maryland School of Public Health

S132: Methods for measurement error problems and their role in improving EHR data-based discovery

Room: R109
Chair: Pamela Shaw, Associate Professor of Biostatistics
Organizer: Yong Chen, University of Pennsylvania

13:30 Methods to address correlated exposure and outcome error for failure time outcomes
Pamela Shaw, Associate Professor of Biostatistics

13:55 Support Vector Machine with Measurement Error
Xin Liu, Shanghai University of Finance and Economics

14:20 Augmented methods in PheWAS studies for pleiotropic effects using biobank data
Yong Chen, University of Pennsylvania

14:45 Multiwave sampling for two-phase designs
Thomas Lumley, University of Auckland

S086: Real World Data and Evidence for Health Care Decision Making

Room: R110
Chair: Zhong Yuan, Janssen Research & Development
Organizer: Wei Shen, Eli Lilly and Company
Co-Organizer: Guohua (James) Pan, Janssen Research & Development

13:30 Precision Health, Real World Data, and Artificial Intelligence Algorithms
Haoda Fu, Eli Lilly and Company

13:55 Real world data, machine learning and causal inference
Jie Chen, Merck Research Laboratory

14:20 Propensity Score Method to Adjust for Confounding in Observational Research: Progression, Challenges, and Opportunities
Zhong Yuan, Janssen Research & Development

14:45 Multistate modeling and simulation of patient trajectories after allogeneic hematopoietic stem cell transplantation to inform drug development
Jiawei Wei, Novartis

S146: Adanced Statistical Methods for Microbiome Sequencing Data with Applications to Complex Human Diseases

Room: R201
Chair: Xiaojing Zheng, UNC-CH
Organizer: Di Wu, University of North Carolina at Chapel Hill
Co-Organizer: Xiaojing Zheng, UNC-CH

13:30 Evaluation of Statistical Methods for Differential Expression Analysis in Microbiome Metatranscriptomics Data
Di Wu, University of North Carolina at Chapel Hill

14:05 Microbial group association test based on the higher criticism
Ni Zhao, Johns Hopkins University

14:40 Multi-SNP mediation intersection-union test
Xiaojing Zheng, UNC-CH

S165: New Statistical Challenges in Biomedical Research

Room: R202
Chair: Yichuan Zhao, Georgia State University
Organizer: Yichuan Zhao, Georgia State University

13:30 Weighted multiple-quantile classifiers for functional data with application in multiple sclerosis screening
Catherine Liu, The Hong Kong Polytechnic University

13:55 Analysis of semi-competing risks data using Archimedean copula models
Antai Wang, New Jersey Institute of Technology

14:20 Two-way partial AUC and its properties
Hanfang Yang, Renmin University of China

14:45 Tracy-Widom law for the largest eigenvalue of sample covariance matrix generated by VARMA
Yangchun Zhang，Harbin Institute of Technology

S095: Nonparametric or semiparametric inference on complicated data

Room: R301
Chair: Niansheng Tang, Yunnan University
Organizer: Niansheng Tang, Yunnan University

13:30 Bayesian Analysis of Semiparametric Hidden Markov Models with Latent Variables
Jingheng Cai, Sun Yat-sen University

14:05 Penalized generalized empirical likelihood with a diverging number of general estimating equations for censored data
Xiaodong Yan, Shandong University

14:40 A Model-averaging method for high-dimensional regre ssion with missing responses at random
Niansheng Tang, Yunnan University

S108: Topics in survival and longitudinal analysis with applications to clinical studies

Room: R302
Chair: Yingchun Zhou, East China Normal University
Organizer: Chao Zhu, Eli Lilly and Company

13:30 Empirical likelihood for additive hazards regression model with case II interval censored failure time data
Chunjie Wang, Changchun University of Technology

14:05 Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression
Hui Zhao, Zhongnan University of Economics and Law

14:40 Functional Mixed Effects model for joint analysis of longitudinal and cross-sectional growth data
Yingchun Zhou, East China Normal University

S062: Some developments on semiparametric regression models and panel data

Room: R303
Chair: Yang Bai, Shanghai University of Finance and Econimics
Organizer: Guoyou Qin, Fudan University

13:30 Powerful Tests for Parent-of-Origin Effects at Quantitative Trait Loci on the X Chromosome
Wing Kam Fung, The University of Hong Kong

14:05 Subgroup Analysis of Linear Model With Measurement Error
Yang Bai, Shanghai University of Finance and Econimics

14:40 Robust Two-Stage Estimation Procedure for Large-dimensional Elliptical Factor Model
Yong He, Shandong University of Finance and Economics

S149: Designs of Modern Clinical Trials

Room: R107
Chair: Bin Nan, University of California, Irvine
Organizer: Bin Nan, University of California, Irvine

13:30 Integrative analysis of high dimensional data under privacy constraints
Molei Liu, Harvard School of Public Health

13:55 Trial Designs for Evaluating Combination HIV Prevention Approaches
Ying Qing Chen, Fred Hutchinson Cancer Research Center

14:20 Lessons Learned from Adaptive RCT Designs
Daniel Gillen, University of California

14:45 Group Sequential Analysis based on RMST
Lu Tian, Stanford University

S116: Finding structures in complex data

Room: R305
Chair: Xinghao Qiao, London School of Economics
Organizer: Chenlei Leng, University of Warwick

13:30 On Consistency and Sparsity for Large-Scale Curve Time Series with Application to Autoregressions
Xinghao Qiao, London School of Economics

14:05 High-dimensional principal component analysis with heterogeneous missingness
Tengyao Wang, London's Global University

14:40 Optimal nonparametric change point detection and localization
Yi Yu, University of Bristol

S092: Dynamic Design of Optimal Treatment Regimes

Room: R306
Chair: Yichi Zhang, University of Rhode Island
Organizer: Shizhe Chen, University of California

13:30 Optimal dynamic treatment regimes using decision lists
Yichi Zhang, University of Rhode Island

13:55 Improved doubly robust estimation in learning optimal individualized treatment rules
Yinghao Pan, University of North Carolina at Charlotte

14:20 Subagging for Inference of the Mean Outcome Under Optimal Treatment Regimes
Chengchun Shi, NC State University

14:45 Online experiment design for mapping large-scale neural circuits
Shizhe Chen, University of California

S133: Statistical advances in accelerating global health and drug development in special population

Room: R307
Chair: Bingzhi Zhang, Sanofi
Organizer: Zhaoling Meng, Gates Medical Research Institute
Co-Organizer: Jeff Barrett, Bill & Melinda Gates Medical Research Institute

13:30 Application of Model Informed Pediatric Extrapolation in Drug Development
Christine Xu, Sanofi

13:55 Janssen's platform trial in evaluating novel compounds in Crohn's Disease
Karen Xia, Johnson and Johnson

14:20 Real World Data Informed Clinical Development via Modeling and Simulation
Zhaoling Meng, Gates Medical Research Institute

14:45 Discussant
Jun Wang, CFDA

December 21 15:40-17:20

S018: Statistical Methods for Integrative Analysis of Big Biomedical Data

Room: R101
Chair: Qi Long, University of Pennsylvania
Organizer: Qi Long, University of Pennsylvania

15:40 Bayesian Nonparametric Clustering Analysis with an Incorporation of Biological Network for High-Dimensional Multi-Scale Molecular Data
Yize Zhao, Yale University

16:05 Statistical methods for integrative clustering analysis of multi-omics data
Qianxing Mo, Moffitt Cancer Center

16:30 Maintaining Data Privacy in Asynchronous Decentralized Data Fusion
Wenxuan Zhong, University of Georgia

16:55 DeepBiome: a phylogenetic tree regularized deep neural network for microbiome data analysis
Jin Zhou, University of Arizona

S031: Optimization Method and Theory for Big Data

Room: R102
Chair: Anru Zhang, University of Wisconsin-Madison
Organizer: Anru Zhang, University of Wisconsin-Madison

15:40 Phase Transition of Landscape From Narrow to Wide Neural Networks
Ruoyu Sun, University of Illinois at Urbana-Champaign

16:15 Inference and Uncertainty Quantification for Noisy Matrix Completion
Yuxin Chen, Princeton University

16:50 On the Equivalence of Inexact Proximal ALM and A DMM for a Class of Convex Composite Programming
Xudong Li, Fudan University

S043: SIBS Invited Session: Recent Advancement in Biostatistics Methodology

Room: R103
Chair: Chen Hu，Johns Hopkins University
Organizer: Mei-Cheng Wang, Department of Biostatistics, Johns Hopkins University

15:40 Hierarchical Bayesian Spatio-Temporal Models with Application to Birds Population Spread
Xuejing Meng, 1. Simon Fraser University 2. Hubei University of Economics

16:05 Inferring Longitudinal antiretroviral drugs effects on mental health in people with HIV
Yanxun Xu, Johns Hopkins University

16:30 Use of Multistate Model for Multiple Endpoints in Oncology Clinical Trials Analysis and Designs
Chen Hu, Johns Hopkins University

16:55 An ADMM Algorithm for Distributed Sparse Optimal Scoring Classification
Yuanshan Wu, Zhongnan University of Economics and Law

S128: Real world evidence in medical research: methods and applications

Room: R104
Chair: Gang Li, Janssen
Organizer: Xiwu Lin, Johnson & Johnson

15:40 Targeted integrative learning with applications in suicide risk prediction
Kun Chen, University of Connecticut

16:05 Analysis of semi-competing risks data using Archimedean copula models
Antai Wang, New Jersey Institute of Technology

16:30 Perspective Plan for Studies Combining Real World Data Sources
Gang Li, Janssen

16:55 Two-Stage Multi-Factor Adaptive Clinical Trials
Samuel Wu, University of Florida

S009: Theoretical challenges for estimations and predictions for large-scale data

Room: R105
Chair: Fei Xue, University of Illinois
Organizer: Annie Qu, University of Illinois

15:40 Subgroup Analysis Based on Structured Mixed-effects Models for Longitudinal Data
Juan Shen, Fudan University

16:05 Collaborative bipartite ranking for personalized prediction
Junhui Wang, City University of Hong Kong

16:30 A pairwise Hotelling method for testing high-dimensional mean vectors
Tiejun Tong, Hong Kong Baptist University

16:55 Integrating multi-source block-wise missing data in model selection
Fei Xue, University of Illinois

S060: New approaches and modifications to modern computations

Room: R106
Chair: Yuichi Mori, Okayama University of Science
Organizer: Yuichi Mori, Okayama University of Science

15:40 Ensemble Classification via Sufficient Dimension Reduction
Yingcun Xia, National University of Singapore

16:05 Zero-inflated negative-binomial NMF
Hiroyasu Abe, Kyoto University

16:30 Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage
Johan Lim, Seoul National University

16:55 Generalized interventional approach for causal mediation analysis with causally ordered multiple mediators
Sheng-Hsuan Lin, National Chiao-Tung University, Institute of Statistics

S114: Highlights of Statistica Sinica

Room: R107
Chair: Yazhen Wang, University of Wisconsin-Madison
Organizer: Hans-Georg Müller, University of California, Davis

15:40 A Model-averaging method for high-dimensional regression with missing responses at random
Nianshen Tang, Yunnan University

16:15 Estimation of Sparse Functional Additive Models with Adaptive Group LASSO
Jiguo Cao, Simon Fraser University

16:50 Understanding and Utilizing the Linearity Condition in Dimension Reduction
Masayuki Henmi, The Institute of Statistical Mathematics

S045: Recent Development in Risk Measure and Its Application

Room: R108
Chair: Xia Zhao, Shanghai University of International Business and Economics
Organizer: Xia Zhao, Shanghai University of International Business and Economics

15:40 Conditional Tail-Related Risk Estimation Using Composit e Asymmetric Least Squares and Empirical Likelihood
Yi Zhang, Zhejiang University

16:00 Measuring systemic risk contagion effect of the banking industry in China: A directed network approach
Zisheng Ouyang, Hunan University Of Technology and Business

16:20 The finite-time ruin probability of a discrete-time risk model with subexponential and dependent insurance and financial risks
Shijie Wang, Anhui University

16:40 Robust portfolio with multi-objective optimization model under high-dimensional scenarios
Xia Zhao, Shanghai University of International Business and Economics

S048: Statistics decision in Drug Development
Room: R109
Chair: Lian Liu, Simcere
Organizer: Lian Liu, Simcere

15:40 Controlling the familywise expected loss with a decision-theoretic approach
Xiaolei Xun, Fudan University

16:05 Sample Size Determination Concerning Decision Making in Clinical Trials - Two Case Studies
Julie Ma, Gilead Sciences, Inc

16:30 Gating criteria using Bayesian approach in early phase study
Wenxin Liu, Roche China

16:55 Quantitative decision making in preclinical drug discovery
Xikun Wu, BeiGene

S158: Modern Clinical Trial Design and Analysis Methods
Room: R110
Chair: Qian Wu, Fred Hutch
Organizer: Qian Wu, Fred Hutch

15:40 Adaptive borrowing of information across patient subgroups in a basket trial based on distributional discrepancy
Haiyan Zheng, Newcastle University

16:15 Estimation of survival under dependent truncation
Jing Qian, University of Massachusetts

16:50 How to apply the multilevel modeling in large health care administrative data
Jun Guan, Methodologist

S156: Recent advances in complex biological data modeling
Room: R201
Chair: Jianhua Hu, Columbia University
Organizer: Weining Shen, UC Irvine

15:40 Covariate-dependent graphs with application in cancer genomics
Yang Ni, Texas A&M University

16:05 Integrative analysis of multi-platform data
Jianhua Hu, Columbia University

16:30 Semiparametric Model for Bivariate Survival Data Subject to Biased Sampling
Jin Piao, University of Southern California

16:55 Generalized probabilistic principal component analysis
Weining Shen, UC Irvine

S104: Model Selection and Information Criteria
Room: R202
Chair: Yoshiyuki Ninomiya, The Institute of Statistical Mathematics
Organizer: Hidetoshi Shimodaira, Kyoto University / RIKEN AIP

15:40 A Cp Criterion for Semiparametric Causal Inference
Yoshiyuki Ninomiya, The Institute of Statistical Mathematics

16:05 High-dimensionality-adjusted Consistent Information Criterion in Multivariate Linear Models
Hirokazu Yanagihara, Hiroshima University

16:30 Convergence rate of importance weighted orthogonal greedy algorithm
Shinpei Imoti, Hiroshima University

16:55 Risk-estimation based predictive densities for heterosk edastic hierarchical models
Keisuke Yano, The University of Tokyo

S098: Advances in sufficient dimension reduction and its applications
Room: R301
Chair: Wei Luo, Zhejiang University
Organizer: Wei Luo, Zhejiang University

15:40 A Model - free Conditional Screening Approach via Sufficient Dimension Reduction
Xuerong Wen, Missouri University of Science and Technology

16:15 Simultaneous estimation for semi-parametric multi-index models
Wenbo Wu, University of Texas at San Antonio

16:50 Matching Using Sufficient Dimension Reduction for Causal Inference
Yeying Zhu, University of Waterloo

S107: Recent Advances in Probability Theory and Related Fields
Room: R302
Chair: Zhonggen Su, Zhejiang University
Organizer: Zhonggen Su, Zhejiang University

15:40 On Cramer-von Mises statistic for the spectral distribution of random matrices
Zhigang Bao, Hong Kong University of Science and Technology

16:05 Concentration Inequalities for Point Processes
Hanchao Wang, Shandong University

16:30 Crossing probabilities in 2D critical lattice models
Hao Wu, Tsinghua University

16:55 Gaussian unitary ensembles with pole singularities near the soft edge and a system of coupled Painlevé XXXIV equations
Lun Zhang, School of Mathematical Sciences, Fudan University

S113: Current Challenges in Functional Data Analysis
Room: R303
Chair: Jiguo Cao, Simon Fraser University
Organizer: Hans-Georg Müller, University of California, Davis

15:40 Spatially Dependent Functional Data: Covariance Estimation, Principal Component Analysis, and Kriging
Yehua Li, University of California at Riverside

16:05 Weak Separability Test for Spatial Functional Feilds
Fang Yao, Peking University

16:30 On multiple segmentation of a functional data sequence
Jeng-Min Chiou, Academia Sinica

16:55 Wasserstein Gradients for the Temporal Evolution of Probability Distributions
Yaqing Chen, University of California, Davis

S123: Big Data and Artificial Intelligence in Medicine: a Bright Future

Room: R304
Chair: Jim Li, Pfizer Inc
Organizer: Kelly Zou,
Co-Organizer: Jim Li, Pfizer Inc

15:40 Generating Real World Evidence for Non-Communicable Disease Control
Jim Li, Pfizer Inc

16:05 Evaluation of the three-in-one team-based care model on hierarchical diagnosis and treatment patterns among patients with diabetes: a retrospective cohort study using Xiamen's regional electronic health records
Yuji Feng, Beijing Innomed Health and Medical Research Center

16:30 Using big data and artificial intelligence to overcome the challenge of identifying patients with a rare disease without diagnosis code
Jun Su, Bioverative

16:55 Machine Learning and Artificial Intelligence for Healthcare
Haoda Fu, Eli Lilly and Company

S067: High dimensional statistical inference

Room: R305
Chair: Yin Xia, Fudan University
Organizer: Shurong Zheng, Northeast Normal University

15:40 Inter-subject correlation analysis with fMRI data
Hongnan Wang, University of Illinois, Chicago

16:05 Individual Data Protected Integrative Regression Analysis of High-dimensional Heterogeneous Data
Yin Xia, Fudan University

16:55 Focused Generalized Method of Moments for High-Dimensional Causal Structural Learning
Changcheng Li, Penn State University

S117: New methods and theory for analysing Big Data

Room: R306
Chair: Liping Zhu, Renmin University of China
Organizer: Chenlei Leng, University of Warwick

15:40 Least Squares Approximation for a Distributed System
Hansheng Wang, Peking University

16:05 Sparsifying Deep Neural Networks with Generalized Regularized Dual Averaging
Guang Cheng, Purdue Statistics

16:30 Single Index Models for Analysis of Mental Health Data with Functional Covariates
Debajyoti Sinha, Florida State University

16:55 Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression
Liping Zhu, Renmin University of China

S130: The Advances of Powerful Tools for Complex Neuroimaging Data

Room: R307
Chair: Zhenke Wu, University of Michigan
Organizer: Haochang Shou, University of Pennsylvania

15:40 Learning Signal Subgraphs from Longitudinal Brain Networks with Symmetric Bilinear Logistic Regression
Lu Wang, Central South University

16:05 Classifying EEG Functional Connectivity Patterns Using A Multi-Domain Convolutional Neural Network
Chee-Ming Ting, Universiti Teknologi Malaysia

16:30 A functional mixed model for scalar on function regression with application to a functional MRI study
Luo Xiao, North Carolina State University

16:55 Neuroconductor: An R Platform for Medical Imaging Analysis
John Muschelli, Johns Hopkins University

December 22 9:00-10:00

ICSA Keynote Lecture

Room: Crystal Ballroom
Chair：Hongzhe Li, University of Pennsylvania
Organizer： ICSA 2019 Organizing Committee

9:00 Statistical models and methods for educational and psychological
Zhiliang Ying, Columbia University

December 22 10:30-12:10

S008: New statistical learning methods for data science problems

Room: R101
Chair: Peng Wang, University of Cincinnati
Organizer: Annie Qu, University of Illinois

10:30 Correlation Tensor Decomposition and Its Application in Spatial Imaging Data
Xiwei Tang, University of Virginia

10:55 Nonparametric interaction selection and screening
Yushen Dong, University of Illinois at Chicago

11:20 A Parsimonious Personalized Dose Finding Model via Dimension Reduction
Ruoqing Zhu, University of Illinois

11:45 Repro Sampling Method for Joint Inference of Model Selection and Regression Coefficients in High Dimensional Linear Models
Peng Wang, University of Cincinnati

S004: Modeling and inference for high-dimensional data

Room: R102
Chair: Ming-Yen Cheng, Hong Kong Baptist University
Organizer: Ming-Yen Cheng, Hong Kong Baptist University

10:30 Generalized Additive Model: Theory, Methods and Applications over Thirty Years
Lijian Yang, Tsinghua University

10:55 Cointegration Rank Estimation for High-dimensional Time Series with Breaks
Rongmao Zhang, Zhejiang University

11:20 Combining Smoothing Spline with Conditional Gaussian Graphical Model for Density and Graph Estimation
Yuedong Wang, University of California-Santa Barbara

11:45 Predicting time series with abrupt changes and smooth evolutions
Jie Ding, University of Minnesota

S014: Modelling large-scale data with complex structures

Room: R103
Chair: Junhui Wang, City University of Hong Kong
Organizer: Junhui Wang, City University of Hong Kong

10:30 Multilevel Structure Modeling of Wearable Device Data with Applications in Population Studies
Xinyue Li, City University of Hong Kong

11:05 Robust Reduced Rank Regression in a Distributed Setting
Xiaojun Mao, Fudan University

11:40 Statistical Inferences of Linear Forms for Noisy Matrix Completion
Dong Xia, Hong Kong University of Science and Technology

S019: Innovative statistical methods for hypothesis testing for high-dimensional data

Room: R104
Chair: Ming Wang, Pennsylvania State University
Organizer: Qi Long, University of Pennsylvania

10:30 SLOPE meets AMP: Does SLOPE outperform LASSO?
Zhiqi Bu, University of Pennsylvania

10:55 IPAD: Stable Interpretable Forecasting with Knockoffs Inference
Yoshimasa Uematsu, Tohoku University

11:20 Gradient-based sparse principal component analysis with extensions to online learning
Yixuan Qiu, Carnegie Mellon University

11:45 RELAXING THE ASSUMPTIONS OF KNOCKOFFS BY CONDITIONING
Dongming Huang, Harvard University

S112: CWS Special Invited Session: Recent Advances in Statistical Methods for Genomic Data

Room: R105
Chair: Wendy Lou, University of Toronto
Organizer: Wendy Lou, University of Toronto

10:30 Statistical Inference of Chromatin 3D Structures from DNA Methylation Data
Shili Lin, Ohio State University

10:55 Network hub-node prioritization of gene regulation with intra-network association
Chuhsing Kate Hsiao, National Taiwan University

11:20 Gene-set Integrative Omics Analysis Using Tensor-based Association Tests
Jung-Ying Tzeng, NC State University

11:45 Genetic factors selection for association study with imbalanced case-control samples
Charlotte Wang, Department of Mathematics, Tamkang University

S070: Recent advances on precision medicine and biomarker research

Room: R108
Chair: Yingqi Zhao, Fred Hutchinson Cancer Research Center
Organizer: Yingqi Zhao, Fred Hutchinson Cancer Research Center

10:30 Constructing personalized decision algorithm for mHealth applications
Min Qian, Columbia University

10:55 Improved doubly robust estimation in learning optimal individualized treatment rules
Yingqi Zhao, Fred Hutchinson Cancer Research Center

11:20 Optimizing personalized intervention from the aspect of health economics
Shuai Chen, University of California

11:45 Learning Individualized Treatment Rules from Electronic Health Records
Yuanjia Wang, Columbia University

S126: Novel Bayesian Adaptive Clinical Trial Designs for Immunotherapy and Precision Medicine

Room: R109
Chair: Sammi Tang, Servier Pharmaceuticals
Organizer: Ying Yuan, MD Anderson Cancer Center

10:30 ComPAS: A Novel Bayesian drug combination platform trial

design with adaptive shrinkage for I/O check inhibitors
Sammi Tang, Servier Pharmaceuticals

11:05 Incorporating population pharmacokinetics data for Phase I-II dose-schedule finding
Fangrong Yan, China Pharmaceutical University

11:40 Characteristics of early phase trial designs for immunocology and comparisons of common designs
Yaqian Zhu, University of Pennsylvania

S068: Large dimensional random matrix theory and its applications

Room: R110
Chair: Cheng Wang, Shanghai Jiao Tong University
Organizer: Shurong Zheng, Northeast Normal University

10:30 Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression
Cheng Wang, Shanghai Jiao Tong University

10:55 Beta matrix and testing the equality of two high dimensional covariance matrices
Jiang Hu, Northeast Normal University

11:20 The limits of the distant sample spikes for a high-dimen sional generalized Fisher matrix and its applications
Dandan Jiang, Xi'an Jiaotong University

11:45 Community Detection Based on the L_infty convergence of eigenvectors in DCBM
Yan Liu, School of Mathematics and Statistics, Northeast Normal University

S061: Advanced statistical modeling for complex data

Room: R111
Chair: Ying Chen, University of Washington
Organizer: Ying Chen, University of Washington

10:30 New Tests for Equality of Several Covariance Functions for Functional Data
Jia Guo, Zhejiang University of Technology

10:55 Pairwise-rank-likelihood methods for the semiparametric transformation model
Tao Yu, National University of Singapore

11:20 Adaptive log-linear zero-inflated generalized Poisson autor egressive model with applications to crime counts
Xiaofei Xu, National University Of Singapore

11:45 Community Detection on Social Network with Complex Attributes
Wanjie Wang, National University of Singapore

S094: Statistical Methods in Complex Data Analysis

Room: R201
Chair: Jianxin Yin, Renmin University of China
Organizer: Jianxin Yin, Renmin University of China

10:30 A maximum average power test for large scale time-course data of counts with applications to RNA-Seq analysis.
Wen Zhou, Colorado State University

10:55 Distributed Dual Averaging Variational Inference
Shiyuan He, Renmin University of China

11:20 A nonparametric Bayesian approach to simultaneous subject and cell heterogeneity discovery for single cell RNA-seq data
Xiangyu Luo, Renmin University of China

11:45 Broadcasted Nonparametric Tensor Regression
Kejun He, Renmin University of China

S101: New Advance in Bayesian Approach for Complex Data

Room: R202
Chair: Chenglong Ye, University of Kentucky
Organizer: Guanyu Hu, University of Connecticut

10:30 A score-based two-stage Bayesian network method for detecting causal SNPs
Yue Zhang, Shanghai Jiao Tong University

10:55 High-dimensional posterior consistency for hierarchical non-local priors in regression
Xuan Cao, University of Cincinnati

11:20 Bayesian Spatially Dynamic Variable Selection for Spatial Point Process
Jieying Jiao, University of Connecticut

11:45 Bayesian Spatial Homogeneity Pursuit Regression for Count Value Data
Guanyu Hu, University of Connecticut

S131: Statistical Learning for the Analysis of Large-scale Omics Data

Room: R302
Chair: Wei Chen, University of Pittsburgh
Organizer: Ying Ding, University of Pittsburgh

10:30 DeepHiC: Greatly Enhancing Chromatin Interaction Information Using Deep Learning
Chun Li, Case Western Reserve University

10:55 Measurement Errors in Array-Based DNA Methylation Analysis
Weihua Guan, University of Minnesota

11:20 Statistical learning for analyzing single-cell multi-omics data
Wei Chen, University of Pittsburgh

11:45 A powerful method for the estimation of cancer-driver genes using a weighted iterative zero-truncated negative-binomial regression
Miaoxin Li, Sun Yat-sen University

S089: Dimension reduction with applications

Room: R303
Chair: Henry Horng-Shing Lu, NCTU
Organizer: Henry Horng-Shing Lu, NCTU

10:30 Sufficient dimension reduction via random-partitions for the large-p-small-n problem
Su-Yun Huang, Academia Sinica

10:55 Robust linear discriminant analysis based on gamma-divergence
Ting-Li Chen, Academia Sinica

11:20 An adaptive clustering for curve data
Heng-Hui Lue, Tunghai University

11:45 Online Learning for Multiclass Classification with Applications
Henry Horng-Shing Lu, NCTU

S102: Structure and correlation analysis

Room: R304
Chair: Xueqin Wang, Sun Yat-sen University
Organizer: Xueqin Wang, Sun Yat-sen University

10:30
Yaowu Zhang, Shanghai University of Finance and Economics

11:05 Test of Independence via Categorically Weighted Distance Correlation
Wei Zhong, Xiamen University

11:40 Best Subset Selection in Linear, Logistic and Cox PH Models
Canhong Wen, University of Science and Technology of China

S105: Statistical Theory for Neural Networks and Machine Learning

Room: R305
Chair: Taiji Suzuki, The University of Tokyo
Organizer: Hidetoshi Shimodaira, Kyoto University / RIKEN AIP

10:30 Generalization error of deep learning and its learning dynamics from compression ability point of view
Taiji Suzuki, The University of Tokyo

10:55 Fisher information of deep neural networks with random weights
Ryo Karakida, National Institute of Advanced Industrial Science and Technology (AIST)

11:20 Generalization Analysis for Mechanism of Deep Learning via Nonparametric Statistics
Masaaki Imaizumi, The Institute of Statistical Mathematics

11:45 Statistical Inference with Unnormalized Models
Takafumi Kanamori, Tokyo Institute of Technology

S106: Dependent Data Analysis

Room: R306
Chair: Hui Huang, Sun Yat-sen University
Organizer: Xueqin Wang, Sun Yat-sen University

10:30 Bayesian spatio-temporal modeling of Arctic sea ice extent
Bohai Zhang, Nankai University

11:05 Autologistic network model on binary data for disease progressionstudy
Huiyan Sang, Texas A&M University

11:40 Integrative interaction analysis of multi-omics data
Mengyun Wu, Shanghai University of Finance and Economics

S039: Recent Advances in Statistical Methods for Single-cell Analysis

Room: R307
Chair: Yuchao Jiang, University of North Carolina at Chapel Hill
Organizer: Di Wu, University of North Carolina at Chapel Hill

10:30 Statistical analysis of coupled single-cell RNA-seq and immune profiling data
Hongkai Ji, Johns Hopkins Bloomberg School of Public Health

10:55 Gene expression imputation and clustering with batch effect removal in single-cell RNA-seq analysis by deep learning
Mingyao Li, University of Pennsylvania

11:20 SMNN: Batch Effect Correction for Single-cell RNA-seq data via Supervised Mutual Nearest Neighbor Detection
Yun Li, University of North Carolina

11:45 Power analysis for RNA-seq in single cells
Zhijin Wu, Brown University

December 22 13:30-15:10

S141: Recent Progresses on Dimension Reduction & High Dimensional Data Analysis

Room: R101
Chair: Qian Lin, Tsinghua University
Organizer: Ke Deng, Tsinghua University

13:30 Sparse SIR via Lasso
Qian Lin, Tsinghua University

14:05 Bayesian Sufficient Dimension Reduction via Modeling Joint Distributions
Yingkai Jiang, Tsinghua University

14:40 Multiple influential point detection in high dimensional regression spaces
Junlong Zhao, School of Statistics, Beijing Normal University

S024: New Advances in Big Data Analysis

Room: R102
Chair: Guoqing Diao, George Mason University
Organizer: Guoqing Diao, George Mason University

13:30 Estimation of endogenous treatment effect estimation with high dimensional instrumental variables and double selection
Qingliang Fan, Xiamen University

14:05 Supervised Clustering via an Implicit Network for High Dimensional Data
Anand Vidyashankar, George Mason University

14:40 Conditional Adaptive Bayesian Spectral Analysis of Nonstationary Time Series
Scott A. Bruce, George Mason University

S047: Recent Advances in Analytic Methods for Sequencing and Biobank Data

Room: R103
Chair: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center
Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

13:30 Bayesian Covariate-dependent Gaussian Graphical Model
Yingying Wei, The Chinese University of Hong Kong

13:55 Statistical assessment of depth normalization methods for microRNA sequencing
Jian Zou, University of Pittsburgh

14:20 Estimating the effect of covariates on the correlation between bivariate failure times
Sean Devlin, Memorial Sloan Kettering Cancer Center

14:45 Mediation Analyses of Ultraviolet, Air Pollution, and Structural Variations using the Taiwan Biobank
En-Yu Lai, Institute of Statistical Science, Academia Sinica

S153: Recent Advances in Ultrahigh Dimensional Data

Room: R105
Chair: Wei Zhong, Xiamen University
Organizer: Wei Zhong, Xiamen University

13:30 The Lq-norm learning for ultrahigh-dimensional survival data: an integrative framework
Xuerong Chen, Southwestern University of Finance and Economics

13:55 Testing for Homogeneity of Mean Vectors and Covariance Matrices in High-dimension
Wenwen Guo, School of Mathematical Sciences, Capital Normal University

14:20 Ball Covariance: A Generic Measure of Dependence in Banach Space
Wenliang Pan, Sun Yat-sen university

14:45 A nonparametric test for proportional covariance matrices in large dimension and small samples
Kai Xu, Anhui Normal University

S090: Advances in survival analysis in the era of data science

Room: R106
Chair: Yen-Tsung Huang, Academia Sinica
Organizer: Yen-Tsung Huang, Academia Sinica

13:30 Survival Analysis of Two-Level Hierarchical Clustered Data
Weijing Wang, National Chiao Tung U

13:55 Semiparametric regression analysis for length-biased and interval-censored data with a cure fraction
Chyong-Mei Chen, National Yang-Ming University

14:20 Semiparametric copula-based analysis for treatment effects in the presence of treatment switching
Yi-Hau Chen, Academia Sinica

14:45 A nonparametric approach to semi-competing risks via causal mediation modeling
Yen-Tsung Huang, Academia Sinica

S078: Joint Modeling and Classification Models for Complex Biomedical Data

Room: R107
Chair: Wendy Lou, University of Toronto
Organizer: Wendy Lou, University of Toronto
Co-Organizer: Daniel Jeske, University of California

13:30 Multilevel joint modeling of hospitalization and survival in patients on dialysis
Esra Kurum, University of California

13:55 Construction, Visualization and Application of Neutral Zone Classifiers
Daniel Jeske, University of California

14:20 Misspecification of a dependent variable in the logistic model controlling for the repeated longitudinal measures
Yi Ting Hwang, National Taipei University

14:45 Joint Trajectories with Variable Selection
Wendy Lou, University of Toronto

S166: Challenges and Analysis of Complex Data

Room: R108
Chair: Guofen Yan, University of Virginia
Organizer: Guofen Yan, University of Virginia

13:30 Classified mixed logistic model prediction
Hanmei Sun, Shandong Normal University

13:55 Examining causal effects of treatments for patients with infective endocarditis
Yingwei Peng, Queen's University

14:20 Latent Class Modeling of Longitudinal Biomarkers in Patients with Chronic Kidney Diseases
Wei Yang, University of Pennsylvania

14:45 Health outcomes research with electronic health records: opportunities and challenges
Guofen Yan, University of Virginia

S152: High dimensional analysis and application in biomarker identification

Room: R109
Chair: Chad He, Fred Hutch
Organizer: Yumou Qiu,
Co-Organizer: Qian Wu, Fred Hutch

13:30 Independence Structure Test in Ultra High-Dimensional Data
Jing He, Southwestern University of Finance and Economics

13:55 Prognostic biomarker identification and subgroup analysis using high dimensional inference in CAR-T cell immunotherapy trial
Qian Wu, Fred Hutch

14:20 Threshold-based subgroup testing in logistic regression models
Ying Huang, Fred Hutchinson Cancer Research Center

14:45 Information Enhanced Model Selection for High-Dimensional Gaussian Graphic Model with Application to Metabolomics Data
Jiang Gui, Dartmouth College

S125: Advances in Statistical Analysis of Omics Data in Agriculture

Room: R110
Chair: Peng Liu, Iowa State University
Organizer: Peng Liu, Iowa State University

13:30 An Efficient Statistical Method for Genomic Selection
Min Zhang, Purdue University

13:55 Statistics Improves Effectiveness of Genomic Selection in Plant Breeding
Lan Zhu, Oklahoma State University

14:20 Exploring high-throughput plant phenomics and genomics data
Yumou Qiu, Iowa State University

14:45 Feature Selection for Rhizosphere Microbiome Studies in Presence of Confounding Using Standardization
Peng Liu, Iowa State University

S022: Innovative method development for complex survival problems

Room: R111
Chair: Huijuan Ma, East China Normal University
Organizer: Limin Peng, Emory University

13:30 Quantile Regression Models for the Survival Data with Missing Censoring Indicator
Zhiping Qiu, School of Statistics, Huaqiao University

13:55 Parametric mode regression for bounded data
Xianzheng Huang, University of South Carolina

14:20 Semiparametric regression analysis for composite endpoints subject to component-wise censoring
Guoqing Diao, George Mason University

14:45 Modeling daily and weekly moderate and vigorous physical activity using zero-inflated mixture Poisson distribution
Xiaonan Xue, Albert Einstein College of Medicine

S073: New developments in statistical methods and inference

Room: R201
Chair: Mengyun Wu, Shanghai University of Finance and Economics
Organizer: Shuangge Ma, Yale University

13:30 Estimating equation methods for longitudinal studies when drop-outs depend on outcome and uncensored observation process
Xia Cui, Speaker Affiliation

13:50 Statistical Modeling in Non-invasive prenatal screening
Xiaobo Guo, Department of Statistical Science, School of Mathematics, Sun Yat-Sen University

14:10 Model Confidence Bounds for Variable Selection
Yang Li, Renmin University of China

14:30 Testing of covariate effects under ridge regression for high-dimensional data
Xu Liu, Shanghai University of Finance and Economics

14:50 Bayesian Variable Selection for Linear Regression with Interaction Terms
Wensheng Zhu, Northeast Normal University

S143: Statistical Advances and Challenges in Bioinformatics

Room: R202
Chair: Lin Hou, Tsinghua University
Organizer: Ke Deng, Tsinghua University

13:30 miRACLe: improving the prediction of miRNA-mRNA interactions by a random contact model
Qi Li, Speaker Affiliation

13:55 Quantifying the impact of genetically regulated expression on complex traits and diseases
Can Yang, HKUST

14:20 Statistical Analysis of Somatic Mutations in Cancer Genomes
Wei Sun, Fred Hutch

14:45 Dimension Reduction and Dropout Imputation for Single-Cell RNA Sequencing Data Using Constrained Robust Nonnegative Matrix Factorization
Shuqin Zhang, Fudan University

S063: Recent advances in Bayesian analysis of complex data

Room: R302
Chair: Xin Tong, National University of Singapore
Organizer: Wanjie Wang, National University of Signapore

13:30 Distributed Bayesian Inference for Varying Coefficient Spatiotemporal Models
Cheng Li, National University of Singapore

13:55 Data assimilation from a viewpoint of regularization theory
Shuai Lu, Fudan University

14:20 Exploiting sparse conditional structure in MALA-within-Gibbs
Xin Tong, National University of Singapore

14:45 Accelerating Metropolis-within-Gibbs sampler with localized computations of differential equations
Qiang Liu, National University of Singapore

S020: Methodological Advancement in High Dimensional Data Analysis

Room: R303
Chair: Zhangsheng Yu, Shanghai Jiao Tong University
Organizer: Honglang Wang, Indiana University-Purdue University Indianapolis

13:30 Adaptive-to-model checking for regressions with diverging number of predictors
Falong Tan, Hunan University

13:55 Data-driven selection of the number of jumps in regression curves: consistency and error rate control
Guanghui Wang, Nankai Univeristy

14:20 Linear Regression Model with Image Input
Zhangsheng Yu, Shanghai Jiao Tong University

14:45 High-dimensional expectile regression with a possible change point
Feipeng Zhang, Xi'an Jiaotong University

S038: Treatment Effects and Other Emerging Issues in Biomedical Data Science

Room:: R304
Chair: Ronghui Xu, University of California
Organizer: Ronghui Xu, University of California

13:30 Model-Free Causal Inference in Observational Studies
Ying Zhang, University of Nebraska Medical Center

13:55 Estimating Treatment Effect under Additive Hazards Models with High-dimensional Covariates
Jue Hou, Harvard T.H. Chan School of Public Health

14:20 A stochastic search approach to identify subgroups with treatment benefit or harm
Changyu Shen, Harvard Medical School

14:45 Optimal Design and Analysis in Phase II Basket Like Trials
Fang Liu, Merck

S129: Integrative Analyses for Wearable Sensor Data in Clinical Studies

Room: R305
Chair: Chee-Ming Ting, Universiti Teknologi Malaysia
Organizer: Haochang Shou, University of Pennsylvania

13:30 Dynamic Bayesian Prediction and Calibration using Multivariate Sensor Data Streams
Zhenke Wu, University of Michigan

13:55 Body Posture Recognition Based on the Raw Accelerometry Data
Jaroslaw Harezlak, Indiana University School of Public Health

14:20 Which to use: objective measurement or performance test for physical activity?
Jiawei Bai, Johns Hopkins University

14:45 Functional Marginal Structural Models for Time-varying Confounding of Mood Assessments
Haochang Shou, University of Pennsylvania

S137: Causal inferences in survival and mediation analyses

Room: R306
Chair: Xinran Li, University of Illinois at Urbana-Champaign
Organizer: Xinran Li, University of Illinois at Urbana-Champaign

13:30 Instrumental variable estimation of a Cox marginal structural model with time-varying endogenous treatments
Yifan Cui, University of Pennsylvania

13:55 Debiased Inverse-Variance Weighted Estimator in Two-Sample Summary-Data Mendelian Randomization
Ting Ye, University of Pennsylvania

14:20 TBD
Masataka Taguri, Yokohama City University

14:45 The randomization distribution of the logrank statistic
Xinran Li, University of Illinois at Urbana-Champaign

S013: Analysis of High-Dimensional Data

Room: R307
Chair: Jie Zhou, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Organizer: Liuquan Sun, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

13:30 Sparse Composite Quantile Regression with Ultra-high Dimensional Heterogeneous Data
Lianqiang Qu, Central China Normal University

14:05 ADealing with Sparsity and Efficiency via Bagging-based Algorithm for Spatio-temporal Autoregressions
Shaojun Guo, Renmin University of China

14:40 Improved matrix pooling
Qizhai Li, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

December 22 15:40-17:20

S015: Regression and classification for complex data

Room: R101
Chair: Heng Lian, City University of Hong Kong
Organizer: Heng Lian, City University of Hong Kong

15:40 Classification with imperfect training labels
Timothy Cannings, University of Edinburgh

16:15 Irrational Exuberance: Correcting Bias in Probability Estimates
Peter Radchenko, University of Sydney

16:50 Optimal Poisson Subsampling from Massive Data
HaiYing Wang, University of Connecticut

S023: Advancement of Quantile Regression Methodology for Complex Data

Room: R102
Chair: Zhiping Qiu, School of Statistics, Huaqiao University
Organizer: Limin Peng, Emory University

15:40 Heterogeneous Individual Risk Modeling of Recurrent Events
Huijuan Ma, East China Normal University

16:05 Statistical analysis of stochastic gradient descent
Jinfeng Xu, Hong Kong University

16:30 Locally Homogeneous Accelerated Failure Time Model with Time-Dependent Covariates
Tony Sit, The Chinese University of Hong Kong

16:55 Partially linear additive quantile regression in ultra-high dimension
Ben Sherwood, University of Kansas

S027: Recent Developments in Modeling and Estimation for Network Data

Room: R103
Chair: Tianxi Li, University of Virginia
Organizer: Wenbin Lu, NC State University

15:40 Two-Mode Network Autoregressive Model for Large-Scale Networks
Danyang Huang, Renmin University of China

16:05 Network-based Clustering for Varying Coecient Panel Data Models
Tao Huang, Shanghai University of Finance and Economics

16:30 Network Response Regression
Jingfei Zhang, University of Miami

16:55 Collaborative Spectral Clustering in Attributed Networks
Pengsheng Ji, University of Georgia

S164: New Advances in Complex Data Analysis and the Applications

Room: R104
Chair: Yichuan Zhao, Georgia State University
Organizer: Yichuan Zhao, Georgia State University

15:40 Regularization of High-Dimensional Toeplitz Covariance Structure via Entropy Loss Function
Jianxin Pan, University of Manchester, UK

16:05 Feature screening with censored data
Qihua Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

16:30 Conditional Quantile Random Forest with its application for predicting the risk (Post-Traumatic Stress Disorder) PTSD after experienced an acute coronary syndrome
Huichen Zhu, The Hong Kong University of Science and Technology

16:55 Large-scale Spatial Predictive Modeling with Applications to Ecological Remote Sensing Data
Chengliang Tang, Columbia University

S120: High dimensional Statistics and Probability

Room: R105
Chair: Yukun He, University of Zurich
Organizer: Zijian Guo, Rutgers University

15:40 Refined Cramer type moderate deviation thorems for general self-nomalied sums with applications
Qi-Man Shao, Southern University of Science and Technology

16:05 Tests for principal eigenvalues and eigenvectors
Xinghua Zheng, HKUST

16:30 Single eigenvalue fluctuations of sparse Erdős-Rényi graphs
Yukun He, University of Zurich

16:55 Asymptotic Normality of the Maximum Likelihood Estimators in ANOVA Models
Fengnan Gao, Fudan University

S056: Recent Advances on the Analysis of Failure Time Data

Room: R106
Chair: Hui Zhao, Zhongnan University of Economics and Law
Organizer: Jianguo Sun, University of Missouri

15:40 Penalized Generalized Empirical Likelihood with a Diverging Number of General Estimating Equations for Censored Data
Xingqiu Zhao, The Hong Kong Polytechnic University

16:05 Empirical likelihood for additive hazards regression model with case II interval censored failure time data
Chunjie Wang, Changchun University of Technology

16:30 Semiparametric analysis of the additive hazards model with informatively interval-censored failure time data
Shuying Wang, Changchun University of Technology

16:55 Regression Analysis of Case-cohort Studies in the Presence of Dependent Interval Censoring
Mingyue Du, Jilin University

S053: New Advances in High-Dimensional Data Analysis

Room: R107
Chair: Rongmao Zhang, Zhejiang University
Organizer: Rongmao Zhang, Zhejiang University

15:40 Extreme Quantile Estimation for Single Index Model
Deyuan Li, Fudan University

16:05 Quantiles, Expectiles and Jackknife Model Averaging in Ultra-High Dimensional Regressions
Yundong Tu, Peking University

16:30 Testing Serial Correlation and ARCH Effect of High-Dimensional Time-Series Data
Yaxing Yang, Xiamen University

16:55 Subgroup Analysis of Zero-Inflated Poisson Model with Application to Insurance Data
Kun Chen, Southwestern University Of Finance And Economics

S169: Statistical Methods and Theory for Complex and Large Data

Room: R108
Chair: Jiang Gui, Dartmouth College
Organizer: Jiang Gui, Dartmouth College

15:40 How to apply the multilevel modeling in large health care administrative data
Jun Guan, Methodologist

16:15 Panel Data Models with Potentially Misspecified Unknown Factors
Huanjun Zhu, Xiamen University

16:50 Dynamic Functional Connectivity Change-point Detectionbased on Random Matrix Theory
Jaehee Kim, Duksung Women's University

S040: Survival Analysis and Beyond

Room: R201
Chair: Ronghui Xu, University of California
Organizer: Ronghui Xu, University of California

15:40 Sparse Boosting for High-Dimensional Survival Data with Varying Coefficients
Jialiang Li, National University of Singapore

16:05 Joint modeling of quality of life and survival data in palliative care studies
Zhigang Li, University of Florida

16:30 A spline-based nonparametric analysis for interval-censored bivariate survival data
Yuan Wu, Duke university

16:55 Causal Effects on Birth Defects with Missing by Terathanasia
Andrew Ying, University of California

S096: Statistical inference on missing or censored data

Room: R202
Chair: Guoyou Qin, Fudan University
Organizer: Niansheng Tang, Yunnan University

15:40 Copula-based semiparametric analysis for time series data with detection limits
Yanlin Tang, East China Normal University

16:05 Multiply Robust Subgroup Identification for Longitudinal Data with Dropouts
Guoyou Qin, Fudan University

16:30 Bayesian Generalized Method of Moments Analysis for Complex Surveys
Puying Zhao, Yunnan University

16:55 A Vine Copula Approach for Regression Analysis of Bivariate Current Status Data with Informative Censoring
Huiqiong Li, Yunnan University

S026: New Advances in Statistical and Machine Learning Methods for Optimal Treatment Decision Making

Room: R301
Chair: Min Qian, Columbia University
Organizer: Wenbin Lu, North Carolina State University

15:40 Learning Optimal Treatment Regimes Using Electronic Health Records for T2D Patients
Donglin Zeng, University of North Carolina

16:05 Optimal treatment decision by a combined moderator
Yu Cheng, University of Pittsburg

16:30 Left without being seen: The disappearance of impatient patients, combining current-status, right-censored and left-censored data
Yair Goldberg, Technion - Israel Institute of Technology

16:55 Diagnosis-Group-Specific Transitional Care Program Recommendations for Thirty-Day Rehospitalization Reduction
Menggang Yu, University of Wisconsin-Madison

S052: Recent development of Gaussian approximation and its applications

Room: R302
Chair: Jinyuan Chang, Southwestern University of Finance and Economics
Organizer: Jinyuan Chang, Southwestern University of Finance and Economics

15:40 Randomized incomplete U-statistics in high dimensions
Xiaohui Chen, University of Illinois at Urbana

16:05 A Power One Test for Unit Roots Based on Sample Autocovariances
Guanghui Cheng, Guangzhou Univerisity

16:30 Asymptotic mixed normality of realized covariance in high-dimensions
Yuta Koike, University of Tokyo

16:55 A Power One Test for Unit Roots Based on Sample Autocovariances
Jinyuan Chang, Southwestern University of Finance and Economics

S080: Matrix Estimation and Matrix regression

Room: R303
Chair: Xingdong Feng, Shanghai University of Finance and Economics
Organizer: Huazhen Lin, Southwestern University of Finance and Economics

15:40 Estimation of error variance via ridge regression
Xingdong Feng, Shanghai University of Finance and Economics

16:05 Nonparametric Regression with a Randomly Censored Independent Variable
Lei Huang, Southwest Jiaotong University

16:30 Influence Matrix Analysis
Wei Lan, Southwest University of Finance and Economic

16:55 Supervised cluster analysis of non-Gaussian functional data
Jiakun Jiang, Tsinghua University

S071: New Advances of Adaptive Data Collection

Room: R304
Chair: Feifang Hu, George Washington University
Organizer: Feifang Hu, George Washington University

15:40 Statistical Inference for Covariate-Adaptive Randomization Procedures
Wei Ma, Renmin University of China

16:05 Randomization-Based Inference Following Randomized Clinical Trials
William Rosenberger, George Mason University

16:30 Neyman-Pearson classification: parametrics and sample size requirement
Lucy Xia, The Hong Kong University of Science and Technology

16:55 Response-adaptive design for clinical trials with recurrent events data
Siu Hung Cheung, Southern University of Science and Technology

S074: New Modeling Methods for Time Series of Analysis
Room: R305
Chair: Dehui Wang, Chair Affiliation
Organizer: Dehui Wang, Organizer Affiliation

15:40 Flexible bivariate Poisson integer-valued GARCH model
Fukang Zhu, Jilin University

16:05 Detecting mean increases in zero truncated INAR(1) processes
Cong Li, Jilin University

16:30 Random coefficients self-exciting threshold integer-valued autoregressive processes driven by logistic regression
Kai Yang, Changchun University of Technology

16:55 Threshold negative binomial autoregressive model
Mengya Liu, Jilin University

S083: Novel Complex Data Analysis Methods
Room: R306
Chair: Yang Li, Renmin University of China
Organizer: Shuangge Ma, Yale University

15:40 Regionalization of PM2.5 in Jing-Jin-Ji Area Using Convex Clustering
Hui Huang, Sun Yat-sen University

16:05 Genome-wide Association Testing for Pleiotropic Effects using GWAS Summary Statistics
Zhonghua Liu, The University of Hong Kong

16:30 Selection models for the efficient design of family studies
Yujie Zhong, School of Statistics and Management, Shanghai University of Finance and Economics

16:55 Semiparametric Varying-coefficient Study of Mean Residual Life Models with right-censored and length-biased data
Fangfang Bai, University of International Business and Economics

S050: Novel Statistical Methods for Big Health Data
Room: R307
Chair: Zhengjia Chen, Emory University
Organizer: Zhengjia Chen, Emory University

15:40 Clustering of Multivariate Data with Varying Dimensions
Bin Cheng, Columbia University

16:15 An Adaptive Trial Design to Optimize Dose-Schedule Regimes with Delayed Outcomes
Ruitao Lin, The University of Texas MD Anderson Cancer Center

16:50 Prediction of Alzheimer's disease by integrating local brain-network connectome
Yanming Li, University of Michigan

Abstracts

S001: Complex Medical Data Analysis

Efficient estimation of the Nonparametric Mean and Covariance Functions for Longitudinal and Sparse Functional Data

Ling Zhou
Southwestern University of Finance and Economics
E-mail: zhouling@swufe.edu.cn

Abstract: We consider the estimation of mean and covariance functions for longitudinal and sparse functional data by using the full quasi-likelihood coupling a modification of the local kernel smoothing method. The proposed estimators are shown to be consistent, asymptotically normal, and semiparametrically efficient in terms of their linear functionals. Their superiority to the competitors is further illustrated numerically through simula- tion studies. The method is applied to analyze AIDS study and atmospheric study. Supplementary materials for this article are available online.

Bayesian Piecewise Linear Mixed Models with a Random Change Point

Xiang Liu
University of South Florida
E-mail: Xiang.Liu@epi.usf.edu

Abstract: To study the pathogenesis of autoimmune (Type 1) diabetes (T1D), proposed accelerator and overload hypotheses postulate that overweight and rapid growth speed up both beta cell insufficiency and an increased insulin resistance. A child's growth (weight) trajectory during childhood starts with a phase of fast growth and then a phase of slow growth. An individual's growth pattern is important because it might be associated with the risk for either 1) islet autoimmunity, 2) clinical onset of T1D, or both. Here, we introduce a Bayesian two-phase piecewise linear mixed model, where the "change point" is an individual-level random effect corresponding to the timing connecting the two growth phases. This method is used to estimate the weight trajectories for children from the Environmental Determinants of Diabetes in the Young (TEDDY) study and then assess the association between the random effects (pre-change slope, post-change slope, change point) and the risk of either 1) islet autoimmunity or 2) clinical onset. The pre-change slope (i.e., the growth rate in the phase of fast growth) was significantly associated with the risk of islet autoimmunity.

Fiducial Model Selection

Xinmin Li
School of Mathematics & Statistics, Qingdao University
E-mail: xmli@qdu.edu.cn

Abstract: With the advent of the era of big data, model selection has become one of the hot re-search topics in contemporary statistics. Model selection is mainly divided into two theoretical methods: Bayesian model selection and model selection with adding penalty factor. Whether it is from the probability of the model or the model selected by the penalty factor, some criteria will be adopted to avoid the model. This talk introduces a new method based on Fiducial inference method, and compares with other model selection methods.

High-dimensional Tobit models

Hua Liang
George Washington University
E-mail: hliang@gwu.edu

Abstract: To study variable selection for high dimensional Tobit models, we formulate the Tobit models to a single-index model. We hybrid the group exponential lasso for the linear models and univariate regression for the Tobit models to achieve variable selection with group structures taken into consideration. The procedure is computationally efficient and easily implemented. Finite sample experiments show its promising performance. We also illustrate its utility by analyzing a dataset from an HIV/AIDS study.

S002: Recent Advances in Functional Data Analysis

Additive Regression for Predictors of Various Natures and Hilbertian Responses with Application to Censored and Missing DataByeong Park

Seoul National University
E-mail: bupark@stats.snu.ac.kr

Abstract: In this paper we consider a fully nonparametric additive regression model for responses and predictors of various natures. This includes the case of Hilbertian and incomplete responses (like censored or missing responses), and continuous, ordinal discrete or even nominal discrete predictors. We propose a backfitting technique that estimates this additive model, and establish the existence of the estimator and the convergence of the associated backfitting algorithm under minimal conditions. We also develop a general asymptotic theory for the estimator, which includes even the case where there is no continuous predictor in the model. We verify the practical performance of the proposed estimator in an extensive simulation study, and apply the method to three data sets, containing respectively a compositional response, a functional response and a censored scalar response.

Predictive Functional Linear Models with Semiparametric Single-Index Interactions

Naisyin Wang
University of Michigan
E-mail: nwangaa@umich.edu

Abstract: When building a predictive model using both functional and multivariate predictors, it is often crucial to include the interaction between the two sets of predictors. To overcome the curse of dimensionality, we assume the interaction depends on a nonparametric, single-index structure of the multivariate predictor and reduce the dimensionality of the functional predictor using functional principal component analysis (FPCA).

We fit the model using an iterative procedure by minimizing a local quasi-likelihood using truncated FPCA series. By treating the number of FPCA scores as a tuning parameter and allowing it to diverge to infinity, we show that for a wide range of this truncation number and different bandwidths {used by the nonparametric component in the single-index interaction}, the parametric component of the model is root-n consistent and asymptotically normal. In addition, the overall prediction error is dominated by the estimation of the nonparametric function in the single-index interaction: an outcome that leads to a CV-based procedure to select the tuning parameters. We also show that the prediction error in the functional effect enjoys the {minimax} optimal rate in Cai and Hall(2006). In a crop yield prediction application, we show that our single-index interaction model yields lower prediction error than the conventional functional linear model and other competing nonlinear functional regression models.

Abstracts

A new knn-classifier for functional data with applications

Jin-Ting Zhang
Department of Statistics and Applied Probability
E-mail: stazjt@nus.edu.sg

Abstract: In this talk, we discuss a new knn (k-nearest neighbors) classifier for functional data. For supervised classification of functional data, several classifiers have been proposed in the literature, including the well-known classic knn classifier. The classic knn classifier selects k nearest neighbors around a new observation and determines its class-membership according to a majority vote. A difficulty arises when there are two classes having the same largest number of votes. To overcome this difficulty, we propose a new knn classifier which selects k nearest neighbors around a new observation from each class. The class-membership of the new observation is determined by the minimum average distance or semi-distance between the k nearest neighbors and the new observation. Good performance of the new knn classifier is demonstrated by simulation studies and real data examples.

Hypothesis Testing in Large-scale Functional Linear Regression

Kaijie Xue
Nankai university
E-mail: kaijie@nankai.edu.cn

Abstract: We explore the functional linear regression by focusing on the large-scale scenario that the scalar response is associated with potentially an ultra-large number of functional predictors, leading to a more challenging model framework than the classical case. The emphasis is to establish rigorous procedures for testing general hypothesis on an arbitrary subset of regression coefficient functions. Specifically, we exploit the techniques developed for post-regularization inference, and propose a new test for the large-scale functional linear regression based on a decorrelated score function that separates the primary and nuisance parameters in functional spaces. Likewise, we also devise the corresponding decorrelated Wald and likelihood ratio tests and establish the exact equivalence among these three tests for the model under consideration. The proposed test is shown uniformly convergent to the prescribed significance, and its finite sample performance is illustrated via simulation studies and a dataset arising from the Human Connectome Project for identifying brain regions associated with emotional tasks.

S003: Methodologies for complex survival data

Information Synthesis and Variable Selection Using A Penalized Empirical Likelihood Approach

Ying Sheng
University of California at San Francisco
E-mail: ying.sheng@ucsf.edu

Abstract: The aggregate data from large databases has become increasingly available, producing auxiliary information that can be incorporated to regulate the analysis of smaller-scale studies which collect more comprehensive individual-level information. Properly incorporating such information is anticipated to improve accuracy and efficiency but is technically challenging. Auxiliary information is usually at aggregate level and can be available via different statistical quantities. In this paper, we propose a unified approach to synthesize auxiliary information in the logistic regression analysis of individual-level data with a diverging number of parameters. The proposed approach summarizes various types of auxiliary information via a system of nonlinear population estimating equations and incorporates it in variable selection and model estimation using a novel penalized empirical likelihood method. We further extend the approach to account for potential inconsistency between the auxiliary information and study subjects, and uncertainties in the auxiliary information. The resulting estimator possesses the oracle property and is asymptotically more efficient than the usual penalized maximum likelihood estimator.

Simulation studies show that the proposed approach performs better with higher efficiency for model estimation and higher accuracy for variable selection. We applied the proposed approach to a pediatric kidney transplant study analysis for illustration.

Synthesizing Independent Stagewise Trials for Optimal Dynamic Treatment Regimes

Yuan Chen
Columbia University
E-mail: yc3281@cumc.columbia.edu

Abstract: Dynamic treatment regimes (DTRs) adaptively prescribe treatments based on patient's intermediate responses and evolving health status over multiple treatment stages. Data from sequential multiple assignment randomization trials (SMARTs) are recommended to be used for learning DTRs. However, due to re-randomization of the same patients over multiple treatment stages and a prolonged follow-up period, SMARTs are often difficult to implement and costly to manage, and patient adherence is always a concern in practice. To lessen such practical challenges, we propose an alternative approach to learn optimal DTRs by synthesizing independent trials over different stages. Specifically, at each stage, data from a single randomized trial along with patient's natural medical history and health status in previous stages are used. We use a backward learning method to estimate optimal treatment decisions at a particular stage, where patient's future optimal outcome increment is estimated using data observed from independent trials with future stages' information. Under some conditions, we show that the proposed method yields consistent estimation of the optimal DTRs and we obtain the same learning rates as those from SMARTs. We conduct simulation studies to demonstrate the advantage of the proposed method. Finally, we apply the developed method to learn optimal DTRs by stagewise synthesis of two randomized trials of therapies for major depressive disorder (MDD). The advantage of the proposed synthesis is validated on an independent trial of MDD.

Quantile Residual Life Regression Based on Semi-Competing Risks Data

Jin-Jian Hsieh
Department of Mathematics, NATIONAL Chung Cheng UNIVERSITY
E-mail:jinjian.hsieh@gmail.com

Abstract: This paper investigates the quantile residual life regression based on semi-competing risk data. Because the terminal event time dependently censors the non-terminal event time, the inference on the non-terminal event time is not available without extra assumption. Therefore, we assume that the non-terminal event time and the terminal event time follow an Archimedean copula. Then, we apply the inverse probability weight technique to construct an estimating equation of quantile residual life regression coefficients. But, the estimating equation may not be continuous in coefficients. Thus, we apply the generalized solution approach to

overcome this problem. Since the variance estimation of the proposed estimator is difficult to obtain, we use the bootstrap resampling method to estimate it. From simulations, it shows the performance of the proposed method is well. Finally, we analyze the Bone Marrow Transplant data for illustration.

Dimension reduction in multivariate baseline proportional hazards models

Ming-Yueh Huang
Institute of Statistical Science, Academia Sinica
E-mail: myh0728@stat.sinica.edu.tw

Abstract: In many applications, it is important to summarize the hazard ratio of certain primary exposure variables, while controlling for many other covariates flexibly.

In the literature, a continuously-stratified proportional hazards model has been proposed to extend Cox model and allow fully nonparametric modeling on controlled covariates.

However, when the number of covariates is large, the curse of dimensionality leads to unstable estimation of the primary exposure effect.

To address this issue, we will study partial sufficient dimension reduction for survival data by introducing a nested family of multivariate baseline proportional hazards models.

The model maintains the practically desirable hazard-ratio interpretation of target parameters, while allowing data-adaptive dimension reduction of multi-dimensional covariates to reduce the effect of curse of dimensionality.

Under the proposed model, we characterize the semiparametric efficiency bound and propose an efficient estimator.

The efficiency gain compared to the continuously stratified proportional hazards model is also proved.

S004: Modeling and inference for high-dimensional data

Two-step Sparse Boosting for High-Dimensional Longitudinal Data with Varying Coefficients

Ming-Yen Cheng
Hong Kong Baptist University
E-mail: chengmingyen@hkbu.edu.hk

Abstract: Varying-coefficient models are useful for analyzing longitudinal data measured repeatedly over time. Research on variable selection has a new focus on the analysis of high-dimensional longitudinal data. We propose a novel two-step sparse boosting approach, for varying-coefficient model with longitudinal data to carry out the variable selection and the model-based prediction. In the first step, we use the sparse boosting technique to yield an estimate of the correlation structure and in the second step, we take into account of the within-subject correlation structure and conduct variable selection and estimation by sparse boosting again. Extensive simulation studies are conducted to demonstrate the validity of the two-step sparse boosting method. We further demonstrate the proposed methodology by an empirical analysis of yeast cell cycle gene expression data.

Generalized Additive Model: Theory, Methods and Applications over Thirty Years

Lijian Yang
Tsinghua University
E-mail: yanglijian@tsinghua.edu.cn

Abstract: The 1984 Stanford University Biostatistics Division Technical Report of Hastie & Tibshirani (1984) introduced the concept and term "generalized additive model (GAM)", which has been popularized by Hastie & Tibshirani (1986 Statistical Science), and Hastie & Tibshirani (1990 Chapman and Hall, 15306 Google Scholar citations). GAM has since found wide applications from environmental studies to predictive policing, from credit rating to survival
analysis. I will discuss the statistical theory for GAM, the development of which reflects the diversity of statistics as a discipline over the past 30 years. Some interesting real life applications of GAM are also presented.

Cointegration Rank Estimation for High-dimensional Time Series with Breaks

Rongmao Zhang
Zhejiang University
E-mail: rmzhang@zju.edu.cn

Abstract: A novel and simple-to-use procedure for estimating the cointegration rank of a high-dimensional time series system with possible breaks is proposed in this paper. Based on a similar idea to principal component analysis, a new expression of the cointegrated time series is derived, from which the cointegration rank can be estimated by the number of the eigenvalues of certain non-negative definite matrix. This method is different from that of Zhang, Robinson and Yao (ZRY, 2019), which used the cointegrated time series to recover the cointegration space.

There are several advantages of the new method: (a) the dimension of the cointegrated time series is allowed to vary with the sample size; (b) it is model free; (c) it is robust for a linear trend, that is, the cointegration rank can be identified without detrending; (d) it is simple-to-use and robust against possible breaks in trend.

The cointegration rank can be estimated without testing and estimating the break points a priori.

Asymptotic properties of the proposed methods are investigated when the dimension of the time series increases with the sample size. Illustrations of simulations are also reported.

Combining Smoothing Spline with Conditional Gaussian Graphical Model for Density and Graph Estimation

Yuedong Wang
University of California-Santa Barbara
E-mail: yuedong@pstat.ucsb.edu

Abstract: Multivariate density estimation and graphical models play important roles in statistical learning. The estimated density can be used to construct a graphical model that reveals conditional relationships whereas a graphical structure can be used to build models for density estimation. Our goal is to construct a consolidated framework that can perform both density and graph estimation. Denote Z as the random vector of interest with density function f(z). Splitting Z into two parts, Z=(X,Y) and writing f(z)=f(x)f(y|x) where f(x) is the density function of X and f(y|x) is the conditional density of Y|X=x. We propose a semiparametric framework that models f(x) nonparametrically using a smoothing spline ANOVA (SS ANOVA) model and f(y|x) parametrically using a conditional Gaussian graphical model (cGGM). Combining flexibility of the SS ANOVA model with succinctness of the cGGM, this framework allows us to deal with high-dimensional data without assuming a joint Gaussian distribution. We propose a backfitting estimation procedure for the cGGM with a computationally efficient approach for selection of tuning parameters. We also develop a geometric inference approach for edge selection. We

establish asymptotic convergence properties for both the parameter and density estimation. The performance of the proposed method is evaluated through extensive simulation studies and real data applications.

A CONSTRUCTIVE APPROACH TO L_0-PENALIZED REGRESSION

Yanyan Liu
Wuhan University
E-mail: liuyy@whu.edu.cn

Abstract: "We propose a constructive approach to estimating sparse, highdimensional linear regression models. The approach is a computational algorithm motivated from the KKT conditions for the L_0-penalized least squares solutions. It generates a sequence of solutions iteratively, based on support detection using primal and dual information and root finding. We refer to the algorithm as SDAR for brevity. Under a sparse Riesz condition on the design matrix and certain other conditions, we show that with high probability, the L_2 estimation error of the solution sequence decays exponentially to the minimax error bound in O(log(Rsqurt{J})) iterations, where J is the number of important predictors and R is the relative magnitude of the nonzero target coefficients; and under a mutual coherence condition and certain other conditions, the ‘1 estimation error decays to the optimal error bound in O(log(R)) iterations. Moreover the SDAR solution recovers the oracle least squares estimator within a finite number of iterations with high probability if the sparsity level is known. Computational complexity analysis shows that the cost of SDAR is O(np) per iteration. We also consider an adaptive version of SDAR for use in practical applications where the true sparsity level is unknown. Simulation studies demonstrate that SDAR outperforms Lasso, MCP and two greedy methods in accuracy and efficiency."

S005: Machine Learning Methods in Biomedical Science

Data Integration of Multiple Genome-Wide Association Studies Under Group Homogeneous Structure

Yuan Jiang
Oregon State University
E-mail: yuan.jiang@stat.oregonstate.edu

Abstract: Nowadays, it’s common to have a large collection of datasets or findings from similar scientific studies, with the famous example of multiple genome-wide association studies that are investigating the same human disease. To take advantage of these datasets or findings, statisticians have developed data integration methods to combine either raw data or summary statistics from multiple studies in order to increase statistical power. Most data integration methods to date can only combine compatible studies with the same explanatory variables; they also tend to ignore the grouping structure of the explanatory variables. However, incompatible studies with grouped explanatory variables arise frequently from multiple genome-wide association studies that employ different genotyping platforms. Therefore, we propose a new method called “gMeta” that can integrate incompatible raw data or summary statistics under a new group homoge- neous structure by utilizing group regularization principles. gMeta not only promotes statistical powers by assuming homogeneity among group-level signals but also allows heterogeneous individual-level signals from different studies. Simulation studies illus- trate the advantage of gMeta over separate analysis in terms of its consistency and enhanced statistical power for detecting weak signals. Finally, an integrative analysis of multiple genetic datasets on schizophrenia shows the applicability and efficacy of gMeta when it is applied to genome-wide association studies.

A unified machine learning method of determining the minimal important difference with the linear structure

Jiwei Zhao
State University of New York at Buffalo
E-mail: zhaoj@buffalo.edu

Abstract: The minimal important difference (MID), or the minimal clinically important difference (MCID), the smallest change in a treatment outcome that an individual patient would identify as important and which would indicate a change in the patient 欽檚 management, has been a fundamentally critical concept in personalized medicine and population health for decades. However, most of the currently existing methods of determining the MID are ad hoc, and cannot incorporate the covariate factors emerged dramatically as the use of the electronic health records. In this paper, we propose a principled, unified machine learning framework of estimating the MID. In particular, we focus on the MID with the linear structure primarily due to its easy accessibility and simple interpretability. We consider both the traditional low-dimensional and the practical high-dimensional cases pertaining to the covariate factors. We contrast the difference of both situations theoretically and conduct comprehensive simulation studies to reinforce these theoretical findings. We also apply our method to the study of chondral lesions in knee surgery to demonstrate the usefulness of the proposed approach.

A Bayesian Semi-supervised Approach to Key Phrase Extraction with Only Positive and Unlabeled Data

Sherry Wang
Southern Methodist University
E-mail: swang@smu.edu

Abstract: A set of keyphrases is often used as a brief summary of a document as it provides a good coverage of the content. Since manual assignment is tedious and time-consuming, learning methods based on the rankings of importance scores have been developed for keyphrase extraction especially when there exists for a large database or collection of documents. Supervised learning requires a labeled training set that needs to be obtained by human effort. Unsupervised learning does not require any labeled set. However, it is often the case that a small portion of keyphrases can be easily obtained from the title or abstract. We propose a model-based semi-supervised Bayesian learning method for keyphrase extraction, which utilizes the information from known positive labels to improve the scoring process. Unlike previous methods that are purely algorithm-driven, our approach is probabilistic and allows for assessment of estimation uncertainty besides its improved performance in controlling the false discovery rate.

Deep Learning with Graph Structure in Small Samples

Rui Feng
University of Pennsylvania
E-mail: ruifeng@upenn.edu

Abstract: Recently deep learning methods have demonstrated strong capabilities ofl discovering potential complex patterns for predicting outcomes. They typically need to be trained on a large amount of data. However, biomedical studies often have limited sample sizes but large feature spaces and these features may correlate, interact, and jointly affect an outcome. In this talk, we present a deep learning framework that incorporates the known relationship among variables. In each layer of

learning, we focused on multiple local substructures trimmed from the parent feature nodes. We derived a computational-efficient peeling algorithm where features are decomposed into9 independent components and then summarized for each substructure. Each layer of learning is built upon a gradually reduced tree and the parameters in summary features are optimized through backpropagation. We evaluated the performance of our method through simulations and applied it to two real studies (1) to predict individual response speed using regional cerebral blood flow from functional magnetic resonance imaging data; (2) to predict lung transplantation outcome using gene expression profiles in donors'lungs within several transplant-related pathways. Our method demonstrated improved prediction accuracy in small-scale data, compared to conventional penalized regression, classification trees, and feed forward neural network.

S006: Frontiers in Financial Statistics and Beyond

Max-linear regression models with regularization

Zhengjun Zhang
University of Wisconsin
E-mail: zjz@stat.wisc.edu

Abstract: Motivated by the newly developed max-linear competing copula factor models and max-stable nonlinear time series models, we propose a new class of max-linear regression models to take advantages of easy interpretable features embedded in linear regression models. It can be seen that linear relation is a special case of max-linear relation. We develop an EM algorithm based maximum likelihood estimation procedure. The consistency and asymptotics of the estimators for parameters are proved. To advance max-linear models to deal with high dimensional predictors, we adopt the common strategy of regularization in the high dimensional regression literature. We demonstrate the broad applicability of max-linear models using simulation examples and real applications in econometric and business modeling. The results, in terms of predictability, show a significant improvement compared with solely using regular regression models and other existing machine learning methods. The results enhance our understanding of the relationsthip between the response variable and the predictors, and among the predictors as well.

Factor Modeling for Volatility

Yi Ding
Hong Kong University of Science and Technology
E-mail: ydingai@connect.ust.hk

Abstract: Under a high-frequency and high-dimensional setup, we establish a framework to estimate the factor structure in idiosyncratic volatility, and more importantly, stock volatility. We provide explicit conditions for the consistency of conducting principal component analysis on realized volatilities in identifying the factor structure in volatility. Empirically, we confirm the factor structure in idiosyncratic volatilities of S&P 500 Index constituents. Furthermore, with strong empirical evidence, we propose a simplified single factor model for stock volatility, where volatility is represented by a common volatility factor and a multiplicative lognormal idiosyncratic component. We further utilize the simplified single factor model for volatility forecasting and show that our proposed approach outperforms various benchmark methods.

Estimating Large Efficient Portfolios with Heteroscedastic Returns

MENGMENG AO
Xiamen University
E-mail: mandyaomengmeng@gmail.com

Abstract: In this paper, we propose a method to estimate mean-variance efficient portfolios when asset returns can be heteroscedastic. We prove that under mild assumptions, our estimated portfolio asymptotically achieves the maximum expected return and meanwhile satises the risk constraint, thus approaching mean-variance efficiency in a setting where the asset returns may exhibit heavy-tailedness, heteroscedasticity, and volatility clustering. Simulation and empirical studies strongly support our theoretical results.

Specification Tests for Covariance Structures in High-Dimensional Statistical Models

Cheng Yong Tang
Temple University
E-mail: yongtang@temple.edu

Abstract: We consider testing the specifications of the covariance structures in statistical models for high-dimensional data. In particular, we are interested in developing such tests when the random vectors of interests are not directly observable, and have to be derived from some estimated models. Additionally, the covariance specifications may involve extra nuisance parameters whose estimations are required. In a generic additive model setting, we develop and investigate test statistics based on the maximum discrepancy measure calculated from the residuals. To approximate the distributions of the test statistics under the null hypothesis, new multiplier bootstrap procedures are proposed with necessary adjustments incorporating the model and nuisance parameter estimation errors. Our theoretical development elucidates the impact due to the model and parameter estimation errors in different settings, and establishes the validity of our testing procedures. Extensive simulations and real data examples confirm the results from our analysis, and demonstrate the performance of the specification tests.

S007: Recent Advances on Large Complex Data

Information Based Complexity of High Dimensional Sparse Functions

Ming Yuan
Columbia University
E-mail: ming.yuan@columbia.edu

Abstract: We investigate optimal algorithms for optimizing and approximating a general high dimensional smooth and sparse function from the perspective of information based complexity. Our algorithms and analyses reveal several interesting characteristics for these tasks. In particular, somewhat surprisingly, we show that the optimal sample complexity for optimization or high precision approximation is independent of the ambient dimension. In addition, we show that the benefit of randomization could be substantial for these problems. Our results illustrate the potential value of experiment design for high dimensional problems.

Inference for Case Probability in High-dimensional Logistic Regression

Zijian Guo
Rutgers University
E-mail: zijguo@stat.rutgers.edu

Abstract: Labeling patients in electronic health records (EHRs) with respect to their statuses of having a disease or condition, i.e. case or control statuses, has increasingly relied on prediction models using high-dimensional variables derived from structured and unstructured EHR

data. A major hurdle currently is a lack of valid statistical inference methods for the case probabilities. In this paper, considering high-dimensional logistic regression models for prediction, we propose a bias-corrected estimator for the case probability through integration of linearization and variance enhancement techniques. We establish asymptotic normality and confidence interval construction of the proposed estimator, and propose a hypothesis testing method for patient case-control labelling. The main novelty of our theoretical development is to establish the asymptotic normality of the data-dependent weighted summation of model errors through employing contraction principles, instead of creating independence by sample splitting. This technique can be of independent interest in studying other high-dimensional inference problems in nonlinear models. The validity of our method does not require sparsity conditions on either the loading vector or the precision matrix of the random design. We demonstrate our method via extensive simulation studies and application to a real data set from Penn Medicine EHR.

Statistical learning for individualized asset allocation

Rui Song
North Carolina State University
E-mail: rsong@ncsu.edu

Abstract: We establish a statistical learning framework for individualized asset allocation. A high-dimensional Q-learning methodology is proposed for continuous decision making. The proposed methodology enjoys desirable theoretical properties and facilitates valid statistical inference for optimal values. Empirically, the proposed statistical learning framework is exercised with Health and Retirement Study data. The results show that our proposed optimal individualized strategy improves individual financial well-being and surpasses benchmark strategies under a consumption-based utility framework.

Change-detection-assisted multiple testing for spatiotemporal data

LILUN DU
Hong Kong University of Science and Technology
E-mail: dulilun@ust.hk

Abstract: We consider large-scale multiple testing of data with spatially and temporally clustered signals. When the conventional false discovery rate (FDR) procedure is applied without taking into account the clustering structure, the power to detect statistical significance tends to be reduced. We formulate a spatiotemporal framework in the presence of multiple change points for multiple testing, and propose a data-driven procedure that aims to fully utilize the clustering information. With the aim of grouping data into several sets, we develop a new change-point detection algorithm that integrates the kernel-based aggregation of spatial observations with a global loss function at the temporal level. Then, we derive an FDR control scheme for set-wise multiple testing. Under some mild conditions on the spatiotemporal dependence structure, the FDR is shown to be strongly controlled. Theoretical analysis and numerical studies demonstrate the advantages of our algorithm over competing methods.

S008: New statistical learning methods for data science problems

Correlation Tensor Decomposition and Its Application in Spatial Imaging Data

Xiwei Tang
University of Virginia
E-mail: xt4yj@virginia.edu

Abstract: Most of existing statistical models in imaging analysis only focus on the first moment information of imaging pixels, while the important pixel-wise correlation structure is usually ignored. In this paper, motivated by the multimodal optical imaging data in a breast cancer study, we propose a new tensor learning approach to analyze spatial-correlated imaging data. Specifically, we construct a higher-order correlation tensor which effectively preserves the spatial information and captures the pixel-wise correlation structure. In addition, we propose a new semi-symmetric tensor decomposition method to model spatial correlations, which enables us to identify spatial structures associated with disease, and thus improves the diagnostic power. We also establish the theoretical properties for recovering the true spatial correlation structure, and develop scalable computational algorithm. We illustrate the performance of the proposed method in both simulation studies and the application to multi-photon breast cancer imaging data. The numerical results indicate that the proposed method outperforms other competing methods including the Convolutional Neural Network (CNN), especially when the sample size of imaging data is limited.

Nonparametric interaction selection and screening

Yushen Dong
University of Illinois at Chicago
E-mail: ydong37@uic.edu

Abstract: Variable selection has been well studied in the recent literature due to the surge of enormous high dimensional data. Interaction between predictors are commonly expected to exist in all kinds of real applications. Recently some parametric interaction selection methods have been proposed. In this talk, we will present a new method to perform nonparametric interaction selection and screening.

A Parsimonious Personalized Dose Finding Model via Dimension Reduction

Ruoqing Zhu
University of Illinois Urbana-Champaign
E-mail: rqzhu@illinois.edu

Abstract: Learning an individualized dose rule in personalized medicine is a challenging statistical problem. Existing methods for estimating the optimal individualized dose rule often suffer from the curse of dimensionality, especially when the dose rule is learned nonparametrically using machine learning approaches. To tackle this problem, we propose a dimension reduction framework that effectively reduces the estimation of dose rule in a lower-dimensional subspace of the covariates, leading to a more parsimonious model. To achieve this, the proposed methods exploit that the subspace is spanned by a few linear combinations of the original covariates, which can be solved efficiently using an orthogonality constrained optimization approach. Based on this framework, we propose two approaches: a direct learning approach that yields the dose rule as commonly desired in personalized medicine, and a pseudo-direct learning approach that focuses more on learning the dimension reduction subspace. Under mild regularity assumptions, we show that the estimators of the proposed methods are asymptotically normal. For both approaches, we formulate the numerical optimization problem as solving solutions on the Stiefel manifold. The performances of the proposed methods are evaluated through simulation studies and a warfarin pharmacogenetic dataset.

Repro Sampling Method for Joint Inference of Model Selection

and Regression Coefficients in High Dimensional Linear Models

Peng Wang
University of Cincinnati
E-mail: wangp9@ucmail.uc.edu

Abstract: This paper proposes a new and effective simulation-based approach, called Repro Sampling method, to conduct statistical inference in high dimensional linear models. The Repro method creates and studies the performance of artificial samples (referred to as Repro samples) that are generated by mimicking the sampling mechanism that generated the true observed sample. By doing so, this method provides a new way to quantify model and parameter uncertainty and provide confidence sets with guaranteed coverage rates on a wide range of problems. A general theoretical framework and an effective Monte-Carlo algorithm, with supporting theories, are developed for high dimensional linear models. This method is used to joint create confidence sets of selected models and model coefficients, with both exact and asymptotic inferences are included. It also provides theoretical development to support the computational efficiency. Furthermore, this development allows us to handle inference problems involving covariates that are perfectly correlated. A new and intuitive graphical tool to present uncertainties in model selection and regression parameter estimation is also developed. We provide numerical studies to demonstrate the utility of the proposed method in a range of problems. Numerical comparisons suggest that the method is far better (in terms of improved coverage rates and significantly reduced sizes of confidence sets) than the approaches that are currently used in the literature. The development provides a simple and effective solution for the difficult post-selection inference problems.

S009: Theoretical challenges for estimations and predictions for large-scale data

Subgroup Analysis Based on Structured Mixed-effects Models for Longitudinal Data

Juan Shen
Fudan University
E-mail: shenjuan@fudan.edu.cn

Abstract: In recent years subgroup analysis has emerged as an important tool to identify unknown subgroup memberships. However, subgroup analysis is still under-studied for longitudinal data. In this paper, we propose a structured mixed-effects approach for longitudinal data to model subgroup distribution and identify subgroup membership simultaneously. In the proposed structured mixed-effects model, the heterogeneous treatment effect is modeled as a random effect from a two-component mixture model, while the membership of the mixture model is incorporated using a logistic model with respect to some covariates. One advantage of our approach is that we are able to derive the estimation of the treatment effects through an EM type algorithm which keeps the subgroup membership unchanged over time. Our numerical studies and real data example demonstrate that the proposed model outperforms other competing methods.

Collaborative bipartite ranking for personalized prediction

Junhui Wang
City University of Hong Kong
E-mail: j.h.wang@cityu.edu.hk

Abstract: Personalized prediction arises as an important yet challenging task, which predicts user-specific preferences on a large number of items given limited information. It is often modeled as certain recommender systems focusing on ordinal or continuous ratings. In this talk, I will present a new collaborative ranking system to predict most-preferred items for each user given search queries. Particularly, a -ranker is proposed based on ranking functions incorporating information on users, items, and search queries through latent factor models. Its probabilistic error bound is established showing that its ranking error has a sharp rate of convergence in the general framework of bipartite ranking, even when the dimension of the model parameters diverges with the sample size. Consequently, this result also indicates that the psi-ranker outperforms two major approaches in bipartite ranking: pairwise ranking and scoring. Finally, the proposed psi-ranker is applied to analyze the Expedia dataset with millions of booking records.

A pairwise Hotelling method for testing high-dimensional mean vectors

Tiejun Tong
Hong Kong Baptist University
E-mail: tongt@hkbu.edu.hk

Abstract: For high-dimensional small sample size data, Hotelling's T2 test is not applicable for testing mean vectors due to the singularity problem in the sample covariance matrix. To overcome the problem, there are three main approaches in the literature: replacing the covariance matrix by an identity matrix, refraining from estimating the correlations so that the covariance matrix estimate is a diagonal matrix, and applying regularization methods to obtain an invertible estimate of the covariance matrix. We note, however, that both approaches may have serious limitations and only work well in certain situations. In this paper, we propose a pairwise Hotelling method for testing high-dimensional mean vectors, which, in essence, provides a good compromise between the existing two approaches. To effectively utilize the correlation information, we construct the new test statistics as a summation of Hotelling's test statistics for the covariate pairs with strong correlations and the squared t statistics for the individual covariates with little correlation with others. We further derive the asymptotic null distributions and power functions for the proposed Hotelling tests under certain regularity conditions. Numerical results show that our new tests are able to control the type I error rates, as well as to achieve a higher statistical power compared to existing methods, especially when the covariates are highly correlated. Two real data examples are also analyzed and they both demonstrate the efficacy of our pairwise Hotelling tests.

Integrating multi-source block-wise missing data in model selection

Fei Xue
University of Illinois
E-mail: xuefei012@gmail.com

Abstract: For multi-source data, blocks of variable information from certain sources are likely missing. Existing methods for handling missing data do not take structures of block-wise missing data into consideration. In this paper, we propose a Multiple Block-wise Imputation (MBI) approach, which incorporates imputations based on both complete and incomplete observations. Specifically, for a given missing pattern group, the imputations in MBI incorporate more samples from groups with fewer observed variables in addition to the group with complete observations. We propose to construct estimating equations based on all available information,

and optimally integrate informative estimating functions to achieve efficient estimators. We show that the proposed method has estimation and model selection consistency under both fixed-dimensional and high-dimensional settings. Moreover, the proposed estimator is asymptotically more efficient than the estimator based on a single imputation from complete observations only. In addition, the proposed method is not restricted to missing completely at random. Numerical studies and ADNI data application confirm that the proposed method outperforms existing variable selection methods under various missing mechanisms.

S010: New challenges in nonparametric inference

Nonparametric modeling of heteroscedasticity in multi-dimensional regression

Kyusang Yu
Konkuk University
E-mail: kyusangu@konkuk.ac.kr

Abstract: Generalized additive models provide a way of circumventing curse of dimension in a wide range of nonparametric regression problem. In this talk, we present a multiplicative model for conditional variance functions where one can apply a generalized additive regression method. This approach extends Fan and Yao (1998) to multivariate cases with a multiplicative structure. In this approach, we use squared residuals instead of using logtransformed squared residuals. This idea gives a smaller variance than Yu (2017) when the variance of squared error is smaller than the variance of log-transformed squared error.We provide estimators based on quasi-likelihood and an iterative algorithm based on smooth backfitting for generalized additive models. We also provide some asymptotic properties of estimators and the convergence of proposed algorithm. A numerical study shows the empirical evidence of the theory.

Tail estimation for the spectral density matrix of multivariate Gaussian random fields

Chae Young Lim
Seoul National University
E-mail: limc.stat@gmail.com

Abstract: Multivariate stationary Gaussian random fields are widely used to fit multivariate spatial data. The one to one correspondence between (cross-)covariance functions and (cross-)spectral densities allows us to model (cross-)spectral densities instead of (cross-)covariance functions. In this talk, we consider multivariate stationary Gaussian random fields. Under some assumptions on high-frequency behavior of (cross-)spectral densities, we introduce an approach to estimate parameters that control tail behaviors by minimizing a multivariate local Whittle likelihood type objective function. We show consistency and asymptotic normality of the estimators with simulation results.

Two-step Sparse Boosting for High-Dimensional Longitudinal Data with Varying Coefficients

Ming-Yen Cheng
Hong Kong Baptist University
E-mail: chengmingyen@hkbu.edu.hk

Abstract: Varying-coefficient models are useful for analyzing longitudinal data measured repeatedly over time. Research on variable selection has a new focus on the analysis of high-dimensional longitudinal data. We propose a novel two-step sparse boosting approach, for varying-coefficient model with longitudinal data to carry out the variable selection and the model-based prediction. In the first step, we use the sparse boosting technique to yield an estimate of the correlation structure and in the second step, we take into account of the within-subject correlation structure and conduct variable selection and estimation by sparse boosting again. Extensive simulation studies are conducted to demonstrate the validity of the two-step sparse boosting method. We further demonstrate the proposed methodology by an empirical analysis of yeast cell cycle gene expression data.

S011: Modeling and analysis of spatial point pattern data

Global multivariate point pattern models for rain type occurrence

Mikyoung Jun
Texas A&M University
E-mail: mjun@stat.tamu.edu

Abstract: We seek statistical methods to study the occurrence of multiple rain types observed by satellite on a global scale. The main scientific interests are to relate rainfall occurrence with various atmospheric state variables and to study the dependence between the occurrences of multiple types of rainfall (e.g. short-lived and intense vs. long-lived and weak; the heights of the rain clouds are also considered). Commonly in point process model literature, the spatial domain is assumed to be a small, and thus planar domain. We consider the log-Gaussian Cox Process (LGCP) models on the surface of a sphere and take advantage of cross-covariance models for spatial processes on a global scale to model the stochastic intensity function of the LGCP models. We present analysis results for rainfall observations from the TRMM satellite and atmospheric state variables from MERRA-2 reanalysis data over the tropical Eastern and Western Pacific Ocean, as well as over the entire tropical and subtropical ocean regions. Statistical inference is done through Monte Carlo likelihood approximation for LGCP models. We employ covariance approximation to deal with massive data.

This is joint work with Courtney Schumacher and R. Saravanan.

Spatial Sampling Design using Generalized Neyman-Scott Process

Zhengyuan Zhu
Iowa State University
E-mail: zhuz@iastate.edu

Abstract: In this paper we introduce a new procedure for spatial sampling design. It is found in previous studies (Zhu and Stein 2006) that the optimal sampling design for spatial prediction with estimated parameters is nearly regular with a few clustered points. The pattern is similar to a generalization of the Neyman-Scott (GNS) process (Loh and Yau 2012) which allows for regularity in the parent process. This motivates the use of a realization of the GNS process as sampling design points. This method translates the high dimensional optimization problem of selecting sampling sites into a low dimensional optimization problem of searching for the optimal parameter sets in the GNS process. Simulation studies indicate that the proposed sampling design algorithm is more computationally efficient than traditional methods while achieving similar minimization of the criterion functions. While the traditional methods become computationally infeasible for sample size larger than a hundred, the proposed algorithm is applicable to a size as large as n=1024. A real data example of finding the optimal spatial design for predicting sea surface temperature in the Pacific Ocean is also considered.

This is a joint work with Sze Him Leung, Ji Meng Loh, and Chun Yip Yau

Intensity estimation for spatial point processes

Ottmar Cronie
Umeå University
E-mail: ottmar.cronie@umu.se

Abstract: This talk discusses some recent advances in non-parametric intensity function estimation for point processes. The first part deals with the classical setting of bandwidth selection for kernel estimators. More specifically, it discusses a method which is based on optimality criteria to be minimised, which are derived from the classical Campbell formula for point processes. This new method is fully nonparametric, does not require knowledge of higher-order moments, and is not restricted to a specific class of point process. The second part of the talk discusses a new approach to adaptive intensity estimation. Since adaptive intensity estimators tend to under-smooth the data in certain places, we propose an additional smoothing operation to be applied to such estimators, which is based on resampling the point pattern through independent random thinning. We refer to this operation as "resample-smoothing" and apply it to Voronoi intensity estimators – at a given location the Voronoi intensity estimator is equal to the reciprocal of the size of the Voronoi cell containing that location. Having presented the different methods and some of their properties, we study their performances through simulation studies.

S012: Analysis of Semiparametric Models

Inference in a mixture additive hazards cure model

Haijin HE
Shenzhen University
E-mail: hehj@szu.edu.cn

Abstract: "We propose a mixture additive hazards cure model for survival data with a cure fraction. The proposed model integrates a logistic regression model for the proportion of patients cured of disease and an AH model for the uncured patients. Generalized estimating equations are developed for parameter estimation, and the asymptotic properties of the resulting estimators are established. In addition, model-checking methods are presented to assess the adequacy of the model. The nite-sample performance of the proposed method is evaluated through simulation studies. An application to a human papillomavirus positive oropharyngeal cancer study is conducted to illustrate the proposed method."

Semiparametric Inference for the Functional Cox Model

Meiling Hao
University of International Business and Economics
E-mail: meilinghao@uibe.edu.cn

Abstract: This article studies penalized semiparametric maximum partial likelihood estimation and hypothesis testing for the functional Cox model in analyzing right-censored data with both functional and scalar predictors. Deriving the asymptotic joint distribution of finite-dimensional and infinite-dimensional estimators is a very challenging theoretical problem due to the complexity of semiparametric models. For the problem, we construct the Sobolev space equipped with a special inner product and discover a new joint Bahadur representation of estimators of unknown regression function and coefficients. Using this key tool, we establish the asymptotic joint normality of the proposed estimators and then con-struct a local confidence interval for an unknown slope function. Furthermore, we study a penalized partial likelihood ratio test, show that the test statistic enjoys the Wilks phenomenon, and also verify the optimality of the test. The theoretical results are examined through simulation studies, and a right-censored data example from the Improving Care of Acute Lung Injury Patients study is provided for illustration.

Efficient Fused Learning for Distributed Imbalanced Data

Jie Zhou
Capital Normal University
E-mail: zhoujie@amss.ac.cn

Abstract: Any data set exhibiting an unequal or highly-skewed distribution between its classes/categories can be regarded as imbalanced data.

Due to privacy concern and other technical limitations, imbalanced data distributed across locations/machines cannot be simply combined and stored in a single central location. The common used naive averaging estimate may be unstable for imbalanced data. In this paper, we propose a fused estimation for logistic regression in analyzing distributed imbalanced data by combining all the cases available on all machines, which is stable and efficient. The consistency and asymptotic normality of the proposed estimator are established under regularity conditions. Asymptotic efficiency compared with the oracle estimator based on the entire imbalanced data is also studied. Extensive simulation studies show that the proposed estimator is as efficient as the oracle estimator in various situations.

S013:Analysis of High-Dimensional Data

Sparse Composite Quantile Regression with Ultra-high Dimensional Heterogeneous Data

Lianqiang Qu
Central China Normal University
E-mail: qulianq@mail.ccnu.edu.cn

Abstract: Quantile regression is widely employed in heterogeneous data, but to select covariates that globally affect the response and estimate coefficients simultaneously are very challenging. In this article, we introduce a new globally concerned quantile variable screening method called sparse composite quantile regression (SCQR) for the analysis of ultra-high dimensional heterogeneous data. The proposed method enjoys the sure screening property, can derive a consistent selection path and yields a consistent estimation for coefficients simultaneously across a continuous range of quantile levels. An extended Bayesian information criterion (EBIC) is employed to select the ``best" candidate from the path. Extensive simulation studies demonstrate the usefulness of the proposed method.

Dealing with Sparsity and Efficiency via Bagging-based Algorithm for Spatio-temporal Autoregressions

Shaojun Guo
Renmin University of China
E-mail: sjguo@ruc.edu.cn

Abstract: We consider a class of spatio-temporal models with sparse (autoregressive) coefficient matrices. It extends popular econometric spatial autoregressive panel data models by allowing the neighbourhood influence for each individual (or panel) different from each other. To estimate the model and overcome the innate endogeneity, we propose a class of generalized methods of moment (GMM) estimators to estimate the coefficient matrices. A novel bagging-based estimator is further developed to conquer the over-determined issue which also occurs in Chang et al. (2015) and Dou et al. (2016). An adaptive forward-backward greedy algorithm is proposed to learn the sparse structure of the coefficient matrices. A new BIC-type selection criteria is further developed to conduct

variable selection for GMM estimators. Asymptotic properties are also studied. The proposed methodology is illustrated with extensive simulation studies. A social network dataset is analyzed for illustration purpose.

Improved matrix pooling

Qizhai Li
Academy of Mathematics and Systems Science, CAS
E-mail: liqz@amss.ac.cn

Abstract: Pooled testing is useful to identify positive specimens for large-scale screening. Matrix pooling is one of the commonly used algorithms. In this work, we investigate the properties of matrix pooling and reveal that the efficiency of matrix pooling is related with the magnitude of overlapping among groups. Based on this property, we develop a new design to further improve the efficiency while taking into account of testing error. The efficiency, pooling sensitivity and specificity of this algorithm are explicitly derived and verified through plasmode simulation of detecting acute human immunodeficiency virus among patients who were suspected to have malaria in rural Ugandan. We show that the new design outperforms matrix pooling in efficiency while retain the pooling sensitivity and specificity

S014: Modelling large-scale data with complex structures

Multilevel Structure Modeling of Wearable Device Data with Applications in Population Studies

Xinyue Li
City University of Hong Kong
E-mail: xinyueli@cityu.edu.hk

Abstract: With the recent development and popularity of wearable devices, actigraphy has been widely used in large-scale population studies to provide continuous and objective activity measures and monitor daily sleep-activity patterns. While actigraphy contains rich information, statistical methods to effectively extract and analyze the information are still lacking. How to effectively analyze time-series physical activity data collected for one week, one month or even longer is challenging and how to account for multilevel structures due to data collection procedures is also crucial for estimating covariate effects. In this talk, I will discuss our proposed method for analyzing actigraphy in population studies. To achieve dimension reduction, we applied functional principal component analysis to characterize main activity patterns. To account for the cluster structure due to multistage sampling as well as the multilevel structure due to within-individual observations across time, we used mixed effects models to capture the multilevel information, estimate group effects, and delineate individual activity profiles. Implementation of our methods in an actigraph dataset from middle school students provides novel insights into adolescent activity patterns. We are able to characterize the heterogeneity in daily activity patterns, estimate covariate effects and further identify the association between physical activity and mental health.

Robust Reduced Rank Regression in a Distributed Setting

Xiaojun Mao
Fudan University
E-mail: maoxj@fudan.edu.cn

Abstract: This talk studies the reduced rank regression problem, which assumes a low-rank structure of the coefficient matrix, together with heavy-tailed noises. To address the heavy-tailed noise, we adopt the quantile loss function instead of a commonly used squared loss. However, the non-smooth quantile loss brings new challenges to both computation and the development of statistical properties, especially when the data is large in size and distributed across different machines. To this end, we first transform the response variable and reformulate the problem into a trace-norm regularized least-squares problem, which greatly facilitates the computation. Based on this formulation, we further develop a distributed algorithm. Theoretically, we establish the convergence rate of the obtained estimator and the theoretical guarantee for rank recovery. The simulation analysis is provided to demonstrate the effectiveness of our method.

Statistical Inferences of Linear Forms for Noisy Matrix Completion

Dong Xia
Hong Kong University of Science and Technology
E-mail: madxia@ust.hk

Abstract: We introduce a flexible framework for making inferences about general linear forms of a large matrix based on noisy observations of a subset of its entries. In particular, under mild regularity conditions, we develop a universal procedure to construct asymptotically normal estimators of its linear forms through double-sample debiasing and low-rank projection whenever an entry-wise consistent estimator of the matrix is available. These estimators allow us to subsequently construct confidence intervals for and test hypotheses about the linear forms. Our proposal was motivated by a careful perturbation analysis of the empirical singular spaces under the noisy matrix completion model which might be of independent interest.

S015: Regression and classification for complex data

Classification with imperfect training labels

Timothy Cannings
University of Edinburgh
E-mail: timothy.cannings@ed.ac.uk

Abstract: We study the effect of imperfect training data labels on the performance of classification methods. In a general setting, where the probability that an observation in the training dataset is mislabelled may depend on both the feature vector and the true label, we bound the excess risk of an arbitrary classifier trained with imperfect labels in terms of its excess risk for predicting a noisy label. This reveals conditions under which a classifier trained with imperfect labels remains consistent for classifying uncorrupted test data points. Furthermore, under stronger conditions, we derive detailed asymptotic properties for the popular k-nearest neighbour ($k$nn), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. One consequence of these results is that the $k$nn and SVM classifiers are robust to imperfect training labels, in the sense that the rate of convergence of the excess risks of these classifiers remains unchanged; in fact, it even turns out that in some cases, imperfect labels may improve the performance of these methods. On the other hand, the LDA classifier is shown to be typically inconsistent in the presence of label noise unless the prior probabilities of each class are equal. Our theoretical results are supported by a simulation study. This is joint work with Yingying Fan and Richard Samworth.

Irrational Exuberance: Correcting Bias in Probability Estimates

Peter Radchenko
University of Sydney
E-mail: peter.radchenko@sydney.edu.au

Abstract: Excess Certainty Adjusted Probability" (ECAP)

Nowadays automated algorithms are routinely used to generate probability estimates, often in real time, for a variety of different events. However, even unbiased probability estimators can provide systematically incorrect estimates, especially in situations where large numbers of probability estimates have been generated. There is a need for an approach to adjust the estimates to account for selection bias. We adopt an empirical Bayes framework and show that a variant of Tweedie's formula can be used to implement the adjustment. Our approach, named "Excess Certainty Adjusted Probability" (ECAP), works particularly well in settings where a large number of probability estimates have been observed. ECAP does not need to make any assumptions about the distribution of the underlying true probabilities. Instead, it relies on estimating the marginal distribution of the corresponding probability estimates, a feasible task in the increasingly common situation where a large number of estimates is observed. We will discuss the theoretical and empirical evidence that the ECAP estimates are generally significantly more accurate than the original ones.

Optimal Poisson Subsampling from Massive Data

HaiYing Wang
University of Connecticut
E-mail: haiying.wang@uconn.edu

Abstract: Nonuniform subsampling methods are effective to reduce computational burden and maintain estimation efficiency for massive data. Existing methods mostly focus on subsampling with replacement due to its high computational efficiency. If the data volume is too large so that nonuniform subsampling probabilities can not be calculated all at once, then subsampling with replacement is infeasible to implement. This paper solve this problem by using Poisson subsampling. We first derive optimal Poisson subsampling probabilities in the context of quasi-likelihood estimation under the A- and L-optimality criteria. For a practically implementable algorithm with approximated optimal subsampling probabilities, we establish the consistency and asymptotic normality of the resultant estimators. To address the situation that the full data are stored at multiple locations, we develop a distributed subsampling framework, in which statistics are computed simultaneously on smaller partitions of the full data. Properties of resultant estimators are investigated in terms of both mean square errors and asymptotic distributions. The proposed strategies are illustrated and evaluated through numerical experiments on simulated and real data sets.

S016: Inference with Complex Data

Jump or Kink: Super-efficiency in Segmented Linear Regression Break-point Estimation

Yining Chen
London School of Economics and Political Science
E-mail: y.chen101@lse.ac.uk

Abstract: We consider the problem of segmented linear regression with the focus on estimating of the location of the break-point(s). Let n be the sample size, we show that the global minimax convergence rate for this problem in terms of the mean absolute error is O(n^{-1/3}). On the other hand, we demonstrate the construction of a super-efficient estimator that achieves the pointwise convergence rate of either O(n^{-1}) or O(n^{-1/2}) for every fixed parameter values, depending on whether the structural change is a jump or a kink. We discuss the implications of this example and the potential remedy. We also illustrate this phenomenon in the multivariate setting.

Model-based Outlier Detection in Multivariate Data with Applications to Detecting Cheating in Tests

Yunxiao Chen
London School of Economics and Political Science
E-mail: y.chen186@lse.ac.uk

Abstract: In this talk, we introduce a statistical framework for the detection of outliers in multivariate data, where outliers are defined as observations (rows) and manifest variables (columns) which deviate from a pre-specified factor model. The outliers are modeled by a factor model component with sparse structures in both the observations and the manifest variables. This problem is motivated by an application to cheating detection in education, where an observation is an examinee and a manifest variable is a test item. The outliers correspond to the cheating examinees and leaked test items, where the cheating examinees have cheated in the exam on the leaked items.

A constrained joint likelihood estimator is proposed that detects row- and column outliers by group truncated L1 constraints. Consistency and rate of convergence are established for this estimator. A DC (difference of convex functions) programming algorithm is developed for the computation of this estimator. The proposed method is applied to an educational testing data set and successfully recovers the cheating examinees and leaked items that have been flagged by the testing program. (This is a joint work with Dr. Xiaoou Li and Mr. Haoran Zhang.)

A computationally efficient approach to the multivariate changepoint problem

Idris Eckley
Lancaster University
E-mail: i.eckley@lancaster.ac.uk

Abstract: "Detecting changepoints within multivariate data sequences is a challenge of increasingly growing importance, due to the increasing prevalence of data collected by systems and sensors. In recent years, several important theoretical and methodological breakthroughs have been proposed. However, the challenge of timely and accurate detection of changepoint locations remains, most notably in the setting where different variates may experience changepoints at different times. This talk will describe current work on this problem, outlining the SUBSET method and demonstrating its utility on both simulated data and data provided by an industrial collaborator.

[This is joint work with Sam Tickle and Paul Fearnhead]"

Online High Dimensional Covariance Change Point Detection

Clifford Lam
London School of Economics and Political Science
E-mail: C.Lam2@lse.ac.uk

Abstract: Detecting changes in the covariance structure in time series data is an important problem which finds applications in, e.g., clustering, identification of disease related genes in bioinformatics, or portfolio and risk management in finance, to name but a few areas. In high dimensional time series data where the dimension can be larger than the sample size, this problem can be extremely difficult. For one, the typical statistic for change detection involves the calculation of sample covariance matrix of two different parts of data, which can be affected by the poor performance of sample covariance matrix under high dimensional setting. More importantly, as far as we know there are no covariance detection methods so far that can have very imbalanced sample sizes for the two parts of data involved,

which unfortunately is exactly the setting for online covariance detection, where one part can have good number of data points but the "newer" part of the data may have only finite number of data points.

We propose a series of statistics which can be powerful in covariance change detection under the online setting, with the dimension of each observation vector grows together with or even faster than the sample size, while the "newer" part of the data under suspicion of change can have only finite sample size. Asymptotic normality of these statistics under both no change and change scenarios are proved and demonstrated, while modifications for variance reduction and power boosting are also illustrated with numerical examples. Incorporation of thee into full change points detection algorithm is also discussed.

S017: Innovative Statistical Methods for Analysis of EHR Data

Electronic Health Record Phenotyping using Anchor-Positive and Unlabeled Patients

Jinbo Chen
University of Pennsylvania
E-mail: jinboche@pennmedicine.upenn.edu

Abstract: Phenotyping patients in electronic health records (EHRs) conventionally relied on algorithms learned from labeled cases and controls. Assigning labels requires manual medical chart review and therefore is labor intensive. We developed a phenotyping method when identification of gold-standard controls is prohibitive so that a validation set is not available. Our method relies on a random subset of cases, which can be specified using an expert-derived anchor variable that has excellent positive predictive value and sensitivity independent of predictors. Adopting a maximum likelihood approach to efficiently leveraging data from the anchor-labeled cases and unlabeled patients to develop logistic regression phenotyping models, we propose novel statistical methods for internally assessing model calibration and predictive performance measures. Upon identification of an anchor variable by clinical experts that is scalable and transferable to different practices, our approach should facilitate development of scalable, transferable, and practice-specific phenotyping models. Through phenotyping two cardiovascular conditions in Penn Medicine EHRs, we demonstrate that our method enables accurate semi-automated EHR phenotyping with minimal manual labeling and therefore is expected to greatly facilitate EHR clinical decision support and research.

Data-driven discovery of medical terms from Chinese electronic health records

Sheng Yu
Tsinghua University
E-mail: syu@tsinghua.edu.cn

Abstract: A comprehensive medical terminology is the basis for mining electronic health records (EHR) and a key infrastructure for medical big data analysis. Chinese medical terminology development is extremely lacked behind compared to English, and severely hampers the development of medical artificial intelligence in China. We propose a method that identifies medical terms from the EHR for the automatic construction of a medical terminology. We treat a sentence as an undirected graph, whose nodes are the characters in the sentences, and whose edge weights represent the connection strength computed with corpus statistics – larger weighs indicate the associated characters are more likely to be in the same term/word. The word segmentation is then achieved with spectral graph partition in an unsupervised manner. After segmentation, a Bi-LSTM classifier is applied to remove incorrectly segmented words and nonmedical words/terms from the output.

Functional clustering methods for extracting features from EHR biomarker history data

Jason Roy
Rutgers School of Public Health
E-mail: jason.roy@rutgers.edu

Abstract: In many modern applications, there is interest in predicting subject-specific functions of a variable over time. For example, we might want to know patient specific trends in a biomarker over time. Modeling is needed If there is measurement error in the variable, or if gaps between data collection times is too wide. We propose a novel semiparametric model for the joint distribution of a continuous longitudinal outcome and the baseline covariates using an enriched Dirichlet process (EDP) prior. This joint model decomposes into subject-specific linear mixed models for the outcome given the covariates and simple marginals for the covariates. The nonparametric EDP prior is placed on the regression and spline coefficients, the error variance, and the parameters governing the predictor space. We predict the outcome at unobserved time points for subjects with data at other time points as well as for completely new subjects with covariates only. We find improved prediction over mixed models with Dirichlet process (DP) priors when there are a large number of covariates. Our method is demonstrated with electronic health records consisting of initiators of second generation antipsychotic medications, which are known to increase the risk of diabetes. We use our model to predict laboratory values indicative of diabetes for each individual and assess incidence of suspected diabetes from the predicted dataset. Our model also serves as a functional clustering algorithm in which subjects are clustered into groups with similar longitudinal trajectories of the outcome over time.

Combining inverse-probability weighting and multiple imputation to adjust for selection bias due to missing data in EHR-based research

Sebastien Haneuse
Harvard T.H. Chan School of Public Health
E-mail: shaneuse@hsph.harvard.edu

Abstract: Among the many potential threats to validity in EHR-based studies, selection bias due to missing data is prominent. Existing missing data methods, however, such as inverse-probability weighting (IPW) and multiple imputation (MI), typically fail to acknowledge the complexity of EHR data. To resolve this, Haneuse et al (2016) proposed to modularize the data provenance into a series of sub-mechanisms, each representing a clinical "decision". Based on this we develop a general and scalable framework for estimation and inference for regression models that permits the use of IPW and/or MI to tackle each of the sub-mechanisms in a single analysis. We refer to this as a "blended analysis strategy". We show that the proposed estimator is consistent and asymptotically Normal, and derive a consistent estimator of the asymptotic variance. Simulations show that naïve use of standard methods may result in bias; that the proposed estimators have good small-sample properties; and, that a bias-variance trade-off may manifest as researchers consider how to handle missing data. The proposed methods are illustrated with data from a multi-site EHR-based study of the effect of bariatric surgery on BMI.

S018: Statistical Methods for Integrative Analysis of Big Biomedical Data

Bayesian Nonparametric Clustering Analysis with an Incorporation of Biological Network for High-Dimensional Multi-Scale Molecular Data

Yize Zhao
Yale University
E-mail: yize.zhao@yale.edu

Abstract: Investigating cancer genome based on multi-type omics data and how it advances personalized medicine is a global medical issue. Though some of the existing clustering methods are capable to character certain degree of concordant and heterogeneity across data types, none of them has incorporated biological network information within and across molecular modalities under cancer subtype discovery. Meanwhile, it is biologically important to identify the core set of biomarkers that are informative to the similarity among samples in each subtype. In this work, with the goal to achieve cancer subtype discovery, we construct a unified clustering model with an incorporation of biological network within and across different molecular data types and simultaneously identifying informative molecular biomarkers for each subtype. Different from existing parametric methods, we adopt a nonparametric approach based on Bayesian Dirichlet process mixture (DPM) models, which is more adaptable to different data types, robust to statistical assumptions and has no constrain on the number of clusters. The performance of the proposed model has been assessed by extensive simulation studies and GCTA.

Statistical methods for integrative clustering analysis of multi-omics data

Qianxing Mo
Department of Biostatistics & Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute
E-mail: Qianxing.Mo@moffitt.org

Abstract: In an effort to identify clinically relevant tumor molecular subtypes and driver omics signatures, we developed integrative clustering methods to jointly model multi-omics data. These methods can model multi-omics data with continuous, count, binary and multi-categorical data types. In the integrative clustering framework, latent variables are used to capture the inherent structure of multi-omics data and variable selection algorithms are used to select important omics features. As a result, the samples can be clustered in the low dimensional latent variable space and driver features contributing to the sample clustering are identified through variable selection. Several multi-omics data sets from large scale genomics studies will be used to illustrate the methods.

A powerful AI tool for CHD screening

Wenxuan Zhong
University of Georgia
E-mail: wenxuan@uga.edu

Abstract: Coronary heart disease (CHD) is a global epidemic that leads to 17.92 million (~1/3) of deaths worldwide in 2016. It is reported by the American College of Cardiology that ischemic heart disease, a late stage of CHD, kills 8.92 million people in 2015 and is ranked No. 1 killer among all diseases. The growing mortality rate of CHD not only causes a significant loss on human resources but also causes many social problems. There is an urgent need for preventive methods to reduce the social burden caused by CHD. However, there is no effective CHD screening methods to date due to the high operational cost, the requirement of expensive and high-maintenance equipment, the need of well trained medical staffs and, most importantly, the potential surgical risk and the radiology side effect on subjects. With the fast development of AI technology, many traditional medical practices are substantially simplified with AI assistance. Unfortunately, most existing AI methods, such as CNN, require extensive computation resources and huge training data as input, which limit the clinical applications of many AI algorithm. In this talk, I will introduce a computational efficient statistical leverage method. I will also illustrate a clinical level AI derivative for CHD screening.

DeepBiome: a phylogenetic tree regularized deep neural network for microbiome data analysis

Jin Zhou
University of Arizona
E-mail: jzhou@email.arizona.edu

Abstract: Evidence linking microbiome and human health is rapidly growing, suggesting that the microbiome profile may serve as a novel predictive biomarker for diseases. Two fundamental challenges of analyzing microbiome data are that bacteria count tables are very sparse and bacteria are classified at a hierarchy of taxonomic levels, ranging from species to phylum. Existing statistical and computational tools often focus on identifying the microbiome association either at the community level or at a specific pre-defined taxonomic level to aggregate rare features. They fail to incorporate hierarchical structure information and cannot learn from the data to aggregate microbiome contribution, therefore, leading to inaccurate selection and prediction.

We present DeepBiome, a deep learning model, to uncover the microbiome-phenotype association network and visualize its path to disease. The proposed DeepBiome takes microbiome abundance data as input and uses the phylogenetic taxonomy to guide the decision of the optimal number of layers and neurons in the deep learning architecture. By doing so, we relieve the computation burden of tuning hyperparameters often encountered in the general deep learning architecture. In addition, we introduce a phylogeny regularized weight decay technique to regularize DeepBiome model and avoid overfitting. DeepBiome analyzes the whole microbiome profile and its path to disease and identifies taxa associated with outcome at each taxonomic level. Our algorithm is applicable to both regression and classification problems in microbiome data analysis. The simulation studies and real data analysis show that DeepBiome is a highly accurate, efficient method and provides deep understanding of complex microbiome-phenotype association.

S019: Innovative statistical methods for hypothesis testing for high-dimensional data

SLOPE is better than LASSO: Estimation and Inference of SLOPE via AMP

Zhiqi Bu
University of Pennsylvania
E-mail: zbu@sas.upenn.edu

Abstract: In high-dimensional problem of reconstructing a sparse signal via the sorted L1 penalized estimation, or SLOPE, we apply the approximate message passing (AMP) to SLOPE problem.

We derive the AMP algorithm and state evolution respectively. We then rigorously prove that AMP solution converges to SLOPE minimization solution as iteration increases. We also use the state evolution for

non-separable functions to asymptotically characterize the SLOPE solution.

As a consequence, AMP and state evolution allow us to conduct inference on the SLOPE solution and demonstrate cases where SLOPE is better than LASSO (which is a special case of SLOPE). Our first result is the trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the SLOPE path. Especially, LASSO is known to suffer from Donoho-Tanner phase transition where TPP may be bounded away from 1. In contrast, SLOPE overcomes such phase transition and one of the path can be nicely characterized as a Mobius transformation. Our second result considers fixed signal prior distribution and constructs SLOPE path that has better TPP, FDP and MSE at the same time.

IPAD: Stable Interpretable Forecasting with Knockoffs Inference

Yoshimasa Uematsu
Tohoku University
E-mail: uematsu0911@gmail.com

Abstract: Interpretability and stability are two important features that are desired in many contemporary big data applications arising in statistics, economics, and finance. While the former is enjoyed to some extent by many existing forecasting approaches, the latter in the sense of controlling the fraction of wrongly discovered features which can enhance greatly the interpretability is still largely underdeveloped. To this end, in this article, we exploit the general framework of model-X knockoffs introduced recently in Candès, Fan, Janson and Lv [(2018), "Panning for Gold: 'model X' Knockoffs for High Dimensional Controlled Variable Selection," Journal of the Royal Statistical Society , Series B, 80, 551–577], which is nonconventional for reproducible large-scale inference in that the framework is completely free of the use of p-values for significance testing, and suggest a new method of intertwined probabilistic factors decoupling (IPAD) for stable interpretable forecasting with knockoffs inference in high-dimensional models. The recipe of the method is constructing the knockoff variables by assuming a latent factor model that is exploited widely in economics and finance for the association structure of covariates. Our method and work are distinct from the existing literature in which we estimate the covariate distribution from data instead of assuming that it is known when constructing the knockoff variables, our procedure does not require any sample splitting, we provide theoretical justifications on the asymptotic false discovery rate control, and the theory for the power analysis is also established. Several simulation examples and the real data analysis further demonstrate that the newly suggested method has appealing finite-sample performance with desired interpretability and stability compared to some popularly used forecasting methods.

Gradient-based sparse principal component analysis with extensions to online learning

Yixuan Qiu
Carnegie Mellon University
E-mail: yixuanq@andrew.cmu.edu

Abstract: Sparse principal component analysis (SPCA) is an important technique for dimensionality reduction of high-dimensional data. However, most existing SPCA algorithms are based on non-convex optimization, which provide little guarantee on the global convergence. SPCA algorithms based on a convex formulation, for example the Fantope projection and selection (FPS) model, overcome this difficulty, but are computationally expensive. In this work we study SPCA based on the convex FPS formulation, and propose a new algorithm that is computationally efficient and applicable to large and high-dimensional data sets. Nonasymptotic and explicit error bounds are derived for both the optimization error and the statistical accuracy, which can be used for testing and inference problems. We also extend our algorithm to online learning problems, where data are obtained in a streaming fashion. The proposed algorithm is applied to high-dimensional genetic data for the detection of functional gene groups.

RELAXING THE ASSUMPTIONS OF KNOCKOFFS BY CONDITIONING

Dongming Huang
Harvard University
E-mail: dhuang01@g.harvard.edu

Abstract: The recent paper Candes et al. (2018) introduced model-X knockoffs, a method for variable selection that provably and non-asymptotically controls the false discovery rate with no restrictions or assumptions on the dimensionality of the data or the conditional distribution of the response given the covariates. The one requirement for the procedure is that the covariate samples are drawn independently and identically from a precisely-known (but arbitrary) distribution. The present paper shows that the exact same guarantees can be made without knowing the covariate distribution fully, but instead knowing it only up to a parametric model. Although this idea is simple, even in Gaussian models conditioning on a sufficient statistic leads to a distribution supported on a set of zero Lebesgue measure, requiring techniques from topological measure theory to establish valid algorithms. We demonstrate how to do this for three models of interest, with simulations showing the new approach remains powerful under the weaker assumptions.

S020: Methodological Advancement in High Dimensional Data Analysis

Adaptive-to-model checking for regressions with diverging number of predictors

Falong Tan
Hunan University
E-mail: falongtan@hnu.edu.cn

Abstract: In this paper, we construct an adaptive-to-model residual-marked empirical process as the base of constructing a goodness-of-fit test for parametric single-index models with diverging number of predictors. To study the relevant asymptotic properties, we first investigate, under the null and alternative hypothesis, the estimation consistency and asymptotically linear representation of the nonlinear least squares estimator for the parameters of interest and then the convergence of the empirical process to a Gaussian process. We prove that under the null hypothesis the convergence of the process holds when the number of predictors diverges to infinity at a certain rate that can be of order, in some cases, o(n^(1/3)/log n) where n is the sample size. The convergence is also studied under the local and global alternative hypothesis. These results are readily applied to other model checking problems. Further, by modifying the approach in the literature to suit the diverging dimension settings, we construct a martingale transformation and then the asymptotic properties of the test statistic are investigated. Numerical studies are conducted to examine the performance of the test.

Data-driven selection of the number of jumps in regression curves: consistency and error rate control

Guanghui Wang
Nankai University
E-mail: ghwang.nk@gmail.com

Abstract: In nonparametric regression with jump discontinuities, one of the major challenges is to determine the number of jumps. Most existing approaches are based on sequential tests or are derived from the model selection viewpoint, which inevitably introduce additional tuning parameters that may not be robust for practical use. We develop a data-adaptive framework with the help of an order-preserved sample-splitting strategy. A cross-validation-based criterion is proposed and its selection consistency is established. More interestingly, the proposed framework allows us to move beyond the point estimation---a new selection procedure with uncertainty quantification is proposed. The key idea is to construct a series of statistics with marginal symmetry property and then to utilize the symmetry for constructing a data-driven threshold to control the false discovery rate. The proposed methodology is computationally efficient, and numerical experiments indicate that it is able to deliver more robust detection results than existing methods in finite samples.

Linear Regression Model with Image Input

Zhangsheng Yu
Shanghai Jiao Tong University
E-mail: yuzhangsheng@sjtu.edu.cn

Abstract: We propose to apply the convolutional neural network for the linear model analysis with both parametric and nonparametric components where the nonparametric component is a projection of images. An estimation procedure is proposed and the asymptotic properties of the parametric coefficients and the convergence of the generalization error were derived. Simulation studies show the performance of the parametric coefficient estimator when CNN is applied is better (smaller bias and variance) compared to other approaches using spline, principle component analysis, or LASSO. We applied this procedure to analyze an Alzheimer disease study with brain image.

High-dimensional expectile regression with a possible change point

Feipeng Zhang
Xi'an Jiaotong University
E-mail: zhangfp108@163.com

Abstract: In this paper, we consider a high-dimensional expectile model with a possible change point due to a covariate threshold. We develop a SCAD-penalized estimator of both regression coefficients and the threshold parameter. The proposed estimator not only selects covariates but also selects a model between linear and threshold regression models. We establish the oracle property of the change point in the sense that its asymptotic distribution is the same as if the unknown active sets of regression coefficients were known. Numerical studies and an application to real data illustrate that the proposed method has satisfactory finite sample performance.

S021: Recent Statistical Advances in Biomedical Research

Asymptotic distribution of the bias corrected LSEs in measurement error linear regression models under long memory

Hira Koul
Michigan State University
E-mail: koul@msu.edu

Abstract: This paper derives the consistency and asymptotic distribution of the bias corrected least squares estimators (LSEs) of the regression parameters in linear regression

models when covariates have measurement error and errors and covariates form mutually independent long memory moving average processes. In the structural ME linear regression model, where the unobservable predicting variables are random, the nature of the asymptotic distribution of suitably standardized BC-LSEs depends on the values of $D_{max} = max {d_X + d_vep, d_X + d_u, d_u + d_vep, 2d_u }, $ where $d_X, d_u,$ and $ d_vep$ are the LM parameters of the covariate, ME and regression error processes, respectively. This limiting distribution is Gaussian when $D_{max} 1/2$. In the former case some consistent estimators of the asymptotic variances of these estimators and a log(n)-consistent estimator of an underlying LM parameter are also provided. They are useful in the construction of the large sample confidence intervals for regression parameters.

In the functional measurement error linear regression models, where the unobservable covariates are non-random, the limiting distribution of the BC-LSEs is always a Gaussian distribution, padetermined by the range of the values of d_vep-d_u.

A Copula Model Approach for Regression Analysis of Informatively Interval-censored Failure Time Data

Jianguo Sun
University of Missouri
E-mail: sunj@missouri.edu

Abstract: "Interval-censored failure time data occur quite often and can have complex structures. Sometime one may face interval-censored data with informative interval censoring and has to deal with even more complicated data structures.
For the situation, one commonly used approach is the frailty-based method.In this talk, we will discuss an alternative approach and present some copula model-based methods for regression analysis of such data."

Personalized Treatment Selection for Joint Optimization of Survival and Other Outcomes

SOMNATH DATTA
University of Florida
E-mail: somnath.datta@ufl.edu

Abstract: We propose a novel method for individualized treatment selection when the treatment response is multivariate which includes a survival component that is subject to right censoring. Since our method covers arbitrary number of treatments and outcome variables, it can be applied to a broad set of models. In addition, the performance measure for each component response can be adjusted depending on the nature of the response. As for example, for a survival component, we might use the difference of mean survivals whereas, for some other clinical covariate, a difference of means may be more suitable. The proposed joint optimization method uses a rank aggregation technique to estimate an ordering of treatments based on ranked lists of treatment performance measures. The method has the flexibility to incorporate patient and clinician preferences to the optimal treatment decision on an individual case basis. An empirical study demonstrates the performance of the proposed method in finite samples. We also present a data analysis using a HIV clinical trial data to show the applicability of the proposed procedure for real data.

Oracally Efficient Estimation and Simultaneous Inference in Partially Linear Single-index Models for Longitudinal Data

Suojin Wang
Texas A&M University
E-mail: sjwang@stat.tamu.edu

Abstract: Oracally efficient estimation and asymptotically accurate simultaneous confidence band (SCB) are established for the nonparametric link function in the partially linear single-index models for longitudinal data. The proposed procedure works for possibly unbalanced longitudinal data under general conditions. The link function estimator is shown to be oracally efficient in the sense that it is asymptotically equivalent in the order of $n^{-1/2}$ to that with all true values of the parameters being known oracally. Furthermore, the asymptotic distribution of the maximal deviation between the estimator and the true link function is provided, and hence an SCB for the link function is constructed. Finite sample simulation studies are carried out which support our asymptotic theory. The proposed SCB is applied to analyze a CD4 data set.

S022:Innovative method development for complex survival problems

Quantile Regression Models for the Survival Data with Missing Censoring Indicator

Zhiping Qiu
School of Statistics, Huaqiao University
E-mail: qzp@hqu.edu.cn

Abstract: The quantile regression model is a flexible and useful approach for analyzing the survival data, which allows the effects of the covariates vary with quantiles. In this paper, we propose a class of quantile regression models for the survival data with missing censoring indicator, which allow the effect of the covariates to be varying or constant. Based on inverse probability weighting, estimating equations imputation and augmented inverse probability weighting tech-nique, three weighted profile estimating equations are proposed and the iterative algorithms that are easily implemented are suggested to solve these profile estimating equations. Asymp-totic properties of the resultant estimators and the resampling-based inference procedures are established. Finally, the finite sample performances of the proposed approaches are investigated in simulation studies and a real data application. The proofs of Theorems are provided in the Appendix.

Parametric mode regression for bounded data

Xianzheng Huang
University of South Carolina
E-mail: huang@stat.sc.edu

Abstract: We propose new parametric frameworks of regression analysis with the conditional mode of a bounded response as the focal point of interest. Covariates effects estimation and prediction based on the maximum likelihood method under two new classes of regression models are demonstrated. We also develop graphical and numerical diagnostic tools to detect various sources of model misspecification. Predictions based on different central tendency measures inferred using various regression models are compared in simulations and real life applications.

Semiparametric regression analysis for composite endpoints subject to component-wise censoring

Guoqing Diao
George Mason University
E-mail: gdiao@gmu.edu

Abstract: Composite endpoints with censored data are commonly used as study outcomes in clinical trials. For example, progression-free survival is a widely used composite endpoint, with disease progression and death as the two components. Progression-free survival time is often defined as the time from randomization to the earlier occurrence of disease progression or death from any cause. The censoring times of the two components could be different for patients not experiencing the endpoint event. Conventional approaches, such as taking the minimum of the censoring times of the two components as the censoring time for progression-free survival time, may suffer from efficiency loss and could produce biased estimates of the treatment effect. We propose a new likelihood-based approach that decomposes the endpoints and models both the progression-free survival time and the time from disease progression to death. The censoring times for different components are distinguished. The approach makes full use of available information and provides a direct and improved estimate of the treatment effect on progression-free survival time.

Simulations demonstrate that the proposed method outperforms several other approaches and is robust against various model misspecifications. An application to a prostate cancer clinical trial is provided.

Modeling daily and weekly moderate and vigorous physical activity using zero-inflated mixture Poisson distribution

Xiaonan Xue

Albert Einstein College of Medicine

E-mail: xiaonan.xue@einstein.yu.edu

Abstract: Large epidemiological studies use recently developed accelerometer devices for continuous and objective monitoring of physical activity. Typically, physical movements are collected in 1-minute epochs and a participant's daily counts of minutes spent in light, moderate and vigorous physical activities are calculated. Because of preponderance of zeros, the daily moderate or higher levels of physical activity data have been modeled using zero-inflated distributions. However, these models do not fully account for variations in daily physical activity and cannot be extended to model weekly physical activity explicitly; while weekly physical activity is often used as an indication of a person's average level. To overcome these limitations, we propose to use a zero-inflated Poisson mixture model for daily physical activity, allowing simultaneous assessments of covariates on daily as well as weekly physical activity. Specifically, a latent variable indicating the likelihood of an active day and the amount of exercise given an active day are modeled respectively by a joint random effects model that incorporates heterogeneity across participants and if needed by an additional random effect to address extra variations in daily physical activity. Maximum likelihood estimation are carried out through Gaussian quadrature technique, which is implemented conveniently in an R package GLMM adaptive. The performance of the methods is examined using simulation studies. These methods are applied to data from the Hispanic Community Health Study/Study of Latinos to examine the difference in daily physical activity between weekday and weekend and the difference in daily and weekly physical activity between BMI groups.

S023:Advancement of Quantile Regression Methodology for Complex Data

Heterogeneous Individual Risk Modeling of Recurrent Events

Huijuan Ma
East China Normal University
E-mail: hjma@fem.ecnu.edu.cn

Abstract: Progression of chronic disease is often manifested by repeated occurrences of disease-related events over time. Delineating the heterogeneity in the risk of such recurrent events can provide valuable scientific insight for guiding customized disease management. In this paper, we present a new dynamic modeling framework, which renders a flexible and robust characterization of individual risk of recurrent event through quantile regression that accounts for both observed covariates and unobservable frailty. The proposed modeling requires no distributional specification of the unobservable frailty, while permitting the exploration of dynamic effects of the observed covariates. We develop estimation and inference procedures for the proposed model through a novel adaptation of the principle of conditional score. The asymptotic properties of the proposed estimator, including the uniform consistency and weak convergence, are established. Extensive simulation studies demonstrate satisfactory finite-sample performance of the proposed method. We illustrate the practical utility of the new method via an application to a diabetes clinical trial that explores the risk patterns of hypoglycemia in Type 2 diabetes patients.

Statistical analysis of stochastic gradient descent

Jinfeng Xu
Hong Kong Univ
E-mail: xujf@hku.hk

Abstract: In many applications involving large dataset or online updating, stochastic gradient descent provides a convenient way to compute parameter estimates and has gained increasing popularity due to its numerical convenience and memory efficiency. While the asymptotic properties of SGD-based estimators have been established decades ago, statistical inference such as interval estimation remains much unexplored. The traditional method such as the bootstrap is not computationally feasible since it requires to repeatedly draw independent samples from the entire dataset. The plug-in method is not applicable when there are no explicit formulas for the covariance matrix of the estimator. In this paper, we propose an inferential procedure for stochastic gradient descent. The proposed method is easy to implement in practice. We establish its theoretical properties for a general class of models that includes generalized linear models and quantile regression models as special cases. The finite-sample performance and numerical utility is evaluated by simulation studies and two real data applications

Locally Homogeneous Accelerated Failure Time Model with Time-Dependent Covariates

Tony Sit
The Chinese University of Hong Kong
E-mail: tonysit@sta.cuhk.edu.hk

Abstract: In this paper, we discuss a generalization of the accelerated failure time model for survival data subject to right censoring, which is independent of the actual lifetime conditional on possibly time-varying covariates. We relax the existing assumption of globally homogeneous conditional quantile on the lifetime distribution to only a specific range of quantile levels. By introducing a class of weighted rank-based estimation procedure, our framework allows a quantile localized inference on the covariate effect with less stringent assumption. Meanwhile, the form of the proposed estimating equations can be viewed as a generalization of its counterpart under the accelerated failure time model with time-varying covariates. Numerical studies demonstrate that the proposed estimator overperforms current alternatives under various settings in terms of smaller empirical bias and standard deviation. A perturbation-based resampling method is also provided to reconcile the asymptotic distribution of the parameter estimates. Finally, consistency and weak convergence of the proposed estimator is established via empirical process theory. This is a joint work with George Chi Wing Chu from Columbia University.

Partially linear additive quantile regression in ultra-high dimension

Ben Sherwood
University of Kansas
E-mail: ben.sherwood@ku.edu

Abstract: Partially linear additive quantile regression allows for estimation of a conditional quantile while allowing some predictors to have an unknown relationship with the response. We consider a penalized estimator that simultaneously estimates the partially linear additive model, while performing variable selection on the linear terms. Rates of convergence and an oracle property are established. In addition, an algorithm for the proposed estimator is provided.

S024:New Advances in Big Data Analysis

Supervised Clustering via an Implicit Network for High Dimensional Data

Anand Vidyashankar
George Mason University
E-mail: avidyash@gmu.edu

Abstract: In high dimensional data analysis, where the number of parameters exceeds the sample size, it is critical to identify features that are significantly associated with the response variable. Also, it is important to detect groups of features, referred to as clusters or hubs, which have similar effects on the response variable. This allows one to provide summarized information about the relationship between the clusters and the response variable. In this presentation, we introduce a new network-based approach for a high dimensional data analysis that addresses these issues. Specifically, we describe a method for constructing an implicit network and describe a new supervised clustering algorithm based on the network-wide metrics. We study the properties of the network-wide metrics and establish theoretical guarantees for the consistency of the supervised clustering algorithm in a high dimensional setting.

Conditional Adaptive Bayesian Spectral Analysis of Nonstationary Time Series

Scott Bruce
George Mason University
E-mail: sbruce7@gmu.edu

Abstract: Many studies of biomedical time series signals aim to measure the association between frequency-domain properties of time series and clinical and behavioral covariates. However, the time-varying dynamics of these associations are largely ignored due to a lack of methods that can assess the changing nature of the relationship through time. This article introduces a method for the simultaneous and automatic analysis of the association between the time-varying power spectrum and covariates, which we refer to as conditional adaptive Bayesian spectrum analysis (CABS). The procedure adaptively partitions the grid of time and covariate values

into an unknown number of approximately stationary blocks and nonparametrically estimates local spectra within blocks through penalized splines. CABS is formulated in a fully Bayesian framework, in which the number and locations of partition points are random, and fit using reversible jump Markov chain Monte Carlo techniques. Estimation and inference averaged over the distribution of partitions allows for the accurate analysis of spectra with both smooth and abrupt changes. The proposed methodology is used to analyze the association between the time-varying spectrum of heart rate variability and self-reported sleep quality in a study of older adults serving as the primary caregiver for their ill spouse.

Estimation of endogenous treatment effect estimation with high dimensional instrumental variables and double selection

Qingliang Fan
Xiamen University
E-mail: michaelqfan@gmail.com

Abstract: This paper proposes a method for endogenous treatment effect estimation using a large number of instruments and double selection (Belloni et al., 2016). In the data-rich environment, it is important to select the control variables in the structural equation. At the same time, the potential endogeneity problem for the treatment variable of interests would not be alleviated even when we put many controls. To address this issue, our method integrates the IV estimation as well as the double selection approach. The asymptotic theory of the IV estimator with double selection is developed, namely, the limiting distribution and the model selection consistency. In an empirical study, we investigate the treatment effect of teacher's attentiveness and student's achievement.

S025: Recent Advances in Statistical Genomics

A sparse clustering algorithm for identifying cluster changes across conditions with applications in single-cell RNA-sequencing data

Jun Li
University of Notre Dame
E-mail: jun.li@nd.edu

Abstract: Clustering analysis, in its traditional setting, identifies groupings of samples from a single population/condition. We consider a different setting when the data available are samples from two different conditions, such as cells before and after drug treatment. Cell types in cell populations change as the condition changes: some cell types die out, new cell types may emerge, and surviving cell types evolve to adapt to the new condition. Using single-cell RNA-sequencing data that measure the gene expression of cells before and after the condition change, we propose an algorithm, SparseDC, which identifies cell types, traces their changes across conditions, and identifies genes which are marker genes for these changes. By solving a unified optimization problem, SparseDC completes all three tasks simultaneously. As a general algorithm that detects shared/distinct clusters for two groups of samples, SparseDC can be applied to problems outside the field of biology.

Statistical Analysis of Spatial Expression Pattern for Spatially Resolved Transcriptomic Studies

Xiang Zhou
University of Michigan
E-mail: xzhousph@umich.edu

Abstract: Recent development of various spatially resolved transcriptomic techniques has enabled gene expression profiling on complex tissues with spatial localization information. Identifying genes that display spatial expression pattern in these studies is an important first step towards characterizing the spatial transcriptomic landscape. Detecting spatially expressed genes requires the development of statistical methods that can properly model spatial count data, provide effective type I error control, have sufficient statistical power, and are computationally efficient. Here, we developed such a method, SPARK. SPARK directly models count data generated from various spatial resolved transcriptomic techniques through generalized linear spatial models. With a new efficient penalized quasi-likelihood based algorithm, SPARK is scalable to data sets with tens of thousands of genes measured on tens of thousands of samples. Importantly, SPARK relies on newly developed statistical formulas for hypothesis testing, producing well-calibrated p-values and yielding high statistical power. We illustrate the benefits of SPARK through extensive simulations and in-depth analysis of four published spatially resolved transcriptomic data sets. In the real data applications, SPARK is up to ten times more powerful than existing approaches. The high power of SPARK allows us to identify new genes and pathways in these data that otherwise cannot be revealed by existing approaches.

Dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder

Jin Gu
Tsinghua University
E-mail: jgu@tsinghua.edu.cn

Abstract: Single cell RNA sequencing (scRNA-seq) is a powerful technique to analyze the transcriptomic heterogeneities in single cell level. It is an important step for studying cell sub-populations and lineages based on scRNA-seq data by finding an effective low-dimensional representation and visualization of the original data. The scRNA-seq data are much noiser than traditional bulk RNA-Seq: in the single cell level, the transcriptional fluctuations are much larger than the average of a cell population and the low amount of RNA transcripts will increase the rate of technical dropout events. In this study, we proposed VASC (deep Variational Autoencoder for scRNA-seq data), a deep multi-layer generative model, for the unsupervised dimension reduction and visualization of scRNA-seq data. It can explicitly model the dropout events and find the nonlinear hierarchical feature representations of the original data. Tested on twenty datasets, VASC shows superior performances in most cases and broader dataset compatibility compared with four state-of-the-art dimension reduction and visualization methods. Also, VASC makes better representations for very rare cell populations in the 2D visualization. Then, for a case study of pre-implantation embryos, VASC successfully re-establishes the cell dynamics and identifies several candidate marker genes associated with the early embryo development.

S026: New Advances in Statistical and Machine Learning Methods for Optimal Treatment Decision Making

Learning Optimal Treatment Regimes Using Electronic Health Records for T2D Patients

Donglin Zeng
University of North Carolina
E-mail: dzeng@email.unc.edu

Abstract: We consider using a large scale EHR data from T2D patients to learn optimal treatments among multiple treatment options such as metformin, insulin or their combinations. We first use an integrated model

for sparsely measured biomarkers to uncover different subgroups that represent patient's heterogeneity in disease progression. Within each subgroup, we adopt inverse probability weighting to adjust potential confounders and use match-learning to estimate optimal treatment strategies. Application to the EHRs from T2D patients in one particular healthcare system shows some interesting findings.

Optimal treatment decision by a combined moderator

Yu Cheng
University of Pittsburgh
E-mail: yucheng@pitt.edu

Abstract: To practice personalized medicine, one may look for all possible qualitative interactions with treatment and prescribe different treatment options based on patients' individual characteristics. Existing trial data and observational studies are often utilized to search for qualitative moderators. However, such exploratory analyses may lead to spurious interactions, especially when a long list of variables are tested for interaction. In this talk, we will discuss an ``optimal'' moderator that is a weighted combination of promising treatment modifiers, where the weights depend on the strengths of interaction in individual contributors. A stabilizing procedure will be used to make weights less dependent on a particular dataset. We will illustrate the methods by searching for personalized optimal treatment option in a smoking cessation study.

Left without being seen: The disappearance of impatient patients, combining current-status, right-censored and left-censored data

Yair Goldberg
Technion - Israel Institute of Technology
E-mail: yairgo@technion.ac.il

Abstract: A dynamic treatment regime is a set of decision rules for how to treat a patient at multiple time points. At each time point, a treatment decision is made depending on the patient's medical history up to that point. We consider the infinite-horizon setting in which the number of decision points is very large. Specifically, we consider long trajectories of patients' measurements recorded over time. At each time point, the decision whether to intervene or not is conditional on whether or not there was a change in the patient's trajectory. We present change-point detection tools and show how to use them in defining dynamic treatment regimes. We demonstrate the utility of the proposed change-point detection for detecting sepsis in preterm infants in the intensive care unit.

Diagnosis-Group-Specific Transitional Care Program Recommendations for Thirty-Day Rehospitalization Reduction

Menggang Yu
University of Wisconsin-Madison
E-mail: meyu@biostat.wisc.edu

Abstract: Thirty-day rehospitalization rate is an important measure reflecting the overall performance of health systems. Recently transitional care programs have been initiated to reduce avoidable rehospitalizations. These programs are behavioral medicine interventions and typically non-drug based. The intervening nurses follow up patients after the hospitalization to manage issues to reduce the risk of rehospitalizations during health care transitions. As rehospitalization is a complex process that depends on many factors, it is unlikely that these interventions are effective for all patients across a diverse population. Therefore we consider individualized treatment rules (ITRs) aimed at maximizing overall treatment effectiveness. We investigate our approach in a setting where patients are divided into two diagnosis related groups, medically complicated and uncomplicated. Such division is crucial for the success of the program because intervening medically complicated patients requires much more efforts from the nurses. In addition the intervention effects can greatly vary between the two groups. Therefore, we consider group-specific recommendation rules that can account for scale differences in treatment effects but allow possible similarity of the estimated ITRs. Computation is realized by morphing our problem into solved forms and a wrapper R package is developed for our proposed treatment recommendation framework. We conducted extensive evaluation through both simulation studies and analysis of a TC program.

S027: Recent Developments in Modeling and Estimation for Network Data

Two-Mode Network Autoregressive Model for Large-Scale Networks

Danyang Huang
Renmin University of China
E-mail: dyhuang89@126.com

Abstract: A two-mode network refers to a network where the nodes are classified into two distinct types, and edges can only exist between nodes of different types. In analysis of two-mode networks, one important objective is to explore the relationship between responses of two types of nodes. To this end, we propose a network autoregressive model for two-mode networks. Different network autocorrelation coefficients are allowed. To estimate the model, a quasi-maximum likelihood estimator is developed with high computational cost. To alleviate the computational burden, a least squares estimator is proposed, which is applicable in large-scale networks. The least squares estimator can be viewed as one particular type of generalized methods of moments estimator. The theoretical properties of both estimators are investigated. The finite sample performances are assessed through simulations and a real data example.

Network-based Clustering for Varying Coecient Panel Data Models

Tao Huang
Shanghai University of Finance and Economics
E-mail: huang.tao@mail.shufe.edu.cn

Abstract: In this talk, we introduce a novel varying-coefficient panel-data model with locally stationary regressors and unknown group structures, wherein the number of groups and the group membership are left unspecified. We develop a triple-localization approach to estimate the unknown subject-specific coefficient functions and then identify the latent group structure via community detection. To improve the e ciency of the first-stage estimator, we further propose a two-stage estimation method that enables the estimator to achieve optimal rates of convergence. In the theoretical part of the paper, we derive the asymptotic theory of the resultant estimators. In the empirical part, we present several simulated examples and a real data to illustrate the finite-sample performance of the proposed method.

Network Response Regression

Jingfei Zhang
University of Miami
E-mail: ezhang@bus.miami.edu

Abstract: Multiple-subject network data are fast emerging in recent years,

where a separate network over a common set of nodes is measured for each individual subject, along with rich subject covariates information. Existing network analysis methods have primarily focused on modeling a single network, and are not directly applicable to multiple-subject networks with subject covariates. In this talk, we introduce a new network response model, where the observed networks are treated as matrix-valued responses, and the individual covariates as predictors. We formulate the parameter estimation as a non-convex optimization problem, and develop an efficient alternating gradient descent algorithm. We establish the non-asymptotic error bound for the actual estimator from our optimization algorithm. Built upon this error bound, we derive the strong consistency for network community recovery, as well as the edge selection consistency. We demonstrate the efficacy of our method through two brain connectivity studies.

Collaborative Spectral Clustering in Attributed Networks

Pengsheng Ji
University of Georgia
E-mail: psji@uga.edu

Abstract: We proposed a novel spectral clustering algorithm for attributed networks, where n nodes split into R non-overlapping communities and each node has a $p-$dimensional meta covariate from various of formats such as text, image, speech etc.. The connectivity matrix $W_{n times n}$ is constructed with the adjacent matrix $A_{n times n}$ and covaraite matrix $X_{n times p}$, and $W = (1-alpha)A + alpha K(X,X')$, where $alpha in [0,1]$ is a tuning parameter and K is a Kernel to measure the covariate similarities. We then perform the classical k-means algorithm on the element-wise ratio matrix of the first K leading eigenvector of W. Theoretical and simulation studies showed the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in imbalanced networks where most community detection algorithms fail.

S028: Statistical and Computational Genomics

Statistical methods for high-resolution chromosome conformation data analysis

Wenxiu Ma
University of California Riverside
E-mail: wenxiu.ma@ucr.edu

Abstract: High-throughput methods based on chromosome conformation capture technologies have enabled us to investigate the three-dimensional (3D) genome organization at an unprecedented resolution. However, high-resolution maps of chromatin interactions require costly, extremely deep sequencing and have been achieved for only a small number of cell lines. Without sufficient sequencing depth, the observed chromatin interaction maps are very sparse and noisy, which imposes great statistical and computational challenges. In this talk, I will present our recent work on enhancing the resolution of chromatin interaction maps and statistical denoising with thresholding.

Efficient algorithms for resampling-based hypothesis testing in genomic data analysis

Hui Jiang
University of Michigan
E-mail: jianghui@umich.edu

Abstract: Resampling-based hypothesis testing procedures such as bootstrapping and permutations tests are widely used in genomic data analysis when the distribution of the test statistic is analytically intractable. However, these test procedures are often computationally intensive, especially when the dataset is large, the desired significant level is very small or there are many tests to perform, all of which are commonly encountered scenarios in modern genomic studies. In this talk, I will discuss several computational methods for accelerating such testing procedures while having theoretical justification and empirical evidence of achieving substantial speedup and high accuracy. The methods will be demonstrated with both simulated and real data experiments in genomics.

Personalized Risk Predictions with Deep Learning Methods in the Presence of Missing and Biased Electronic Health Record and Genomics Data

Hua Zhong
NYU Langone Health
E-mail: judy.zhong@nyumc.org

Abstract: Currently available risk prediction methods are limited in their ability to deal with complex, heterogeneous, and longitudinal data such as that available in electronic health records (EHRs) and genomics studies. Recurrent neural networks (RNNs) have shown significant promise in this context, where (essentially) the model learns a latent representation for a patient state over time, updating the state when new covariate measurements arrive, in a flexible non-linear manner. These models have been shown to be very effective in modeling signals such as speech and text sequences, leading to prediction models that are significantly more accurate than previous non-RNN approaches. However, it is not straightforward how to directly apply a conventional RNN to EHR data involving(a) a significant amount of missing data (many covariates such as lab tests might not be measured during a particular patient visit), and (b) asynchronous measurements(patients show up at varying time-intervals). We propose a patient-specific RNN to learn the time-to-event distributions through flexibly incorporating both missing and asynchronous measurements over time. We demonstrate the efficacy of our approach by applying it to a real-world longitudinal EHR dataset to predict cardiovascular disease (CVD) in patients with type 2 diabetes (T2DM).

A linear mixed model framework to study gene-environment interactions

Hung-Chih Ku
DePaul University
E-mail: hku4@depaul.edu

Abstract: Investigating gene-environment (G×E) interactions is an important step towards the complete understanding of complex human diseases. Regression-based method with linearity assumption is the commonly used approach in detecting G×E interactions. However, the assumption could be violated due to non-linear genetic effects. In this talk, we will perform an approach on testing the G×E interactions, evaluate the performance of simulation studies, and apply the approach to real data.

S029: Statistical machine learning in data science

Accurate and Efficient Machine Learning Methods

Ping Li
Baidu Research USA
E-mail: pingli98@gmail.com

Abstract: In this talk, we present a series of new work on highly efficient machine learning methods which can produce accurate results. In the current trend of practice, it has been common to use sophisticated large

models for (hopefully) achieving good accuracy. Our methods, which are also able to produce accurate results, are different in that they appear simple and are indeed easy to implement/deploy in practice. The author has been working on this line of work for the past decade and will present the most recent results based on collaborations with several statisticians and probabilists.

Penalty Method for Variance Component Selection

Hua Zhou
UCLA
E-mail: huazhou@ucla.edu

Abstract: Variance components models, also known as mixed effects model, are a central theme in statistics. When there are a large number of variance components, one wants to select a subset of those that are associated with response. Existing methods are limited to finding random components at individual level or within one variance component. We propose selection of variance components based on a penalized log-likelihood with adaptive penalty. This is solved by a majorization-minimization (MM) algorithm, which is simple, numerically stable, and globally convergent. Performance of the proposed methodology is evaluated empirically through simulation studies and real data analysis. In theory, we establish a non-asymptotic error bound for the output from the algorithm and characterize the region in which the MM iterates converge to a global optimum of the population likelihood. This result provides a theoretical guideline in terminating MM iterations.

Identifiability of nonparametric mixture models, clustering, and semi-supervised learning

Nikhyl Aragam
University of Chicago
E-mail: naragam@cs.cmu.edu

Abstract: Motivated by problems in data clustering and semi-supervised learning, we establish general conditions under which families of nonparametric mixture models are identifiable by introducing a novel framework for clustering overfitted parametric (i.e. misspecified) mixture models. These conditions generalize existing conditions in the literature, allowing for general nonparametric mixture components. Notably, our results avoid imposing assumptions on the mixture components, and instead impose regularity assumptions on the underlying mixing measure. After a discussion of some statistical aspects of this problem, we will discuss two applications of this framework. First, we extend classical model-based clustering to nonparametric settings and develop a practical algorithm for learning nonparametric mixtures. Second, we analyze the sample complexity of semi-supervised learning (SSL) and introduce new assumptions based on the mismatch between a mixture model learned from unlabeled data and the true mixture model induced by the (unknown) class conditional distributions. Under these assumptions, we establish an Omega(Klog K) labeled sample complexity bound without imposing parametric assumptions, where K is the number of classes. These results suggest that even in nonparametric settings it is possible to learn a near-optimal classifier using only a few labeled samples.

Directed acyclic graphs on network data

Qing Zhou
UCLA Department of Statistics
E-mail: zhou@stat.ucla.edu

Abstract: The traditional directed acyclic graph (DAG) model assumes data are generated independently from the underlying joint distribution defined by the DAG. In many applications, however, individuals are linked via a network and thus the independence assumption does not hold. We propose a novel Gaussian DAG model for network data, where the dependence among individual data points (row-wise covariance) is modeled by an undirected graph. Under this model, we develop a maximum penalized likelihood method to estimate the DAG structure and the row correlation matrix. The algorithm iterates between a decoupled lasso regression step and a graphical lasso step. We show with extensive simulated and real network data, that our algorithm improves the accuracy of DAG structure learning by leveraging the information from the estimated row-wise correlations. Moreover, we demonstrate that the performance of existing DAG learning methods can be substantially improved via de-correlation of network data with the estimated row-wise correlation matrix from our algorithm.

S030: Statistical Inference for High-dimensional Tensor Data

Tensor bandits for online interactive recommendation

Wei Sun
Purdue University
E-mail: sun244@purdue.edu

Abstract: Traditional static recommendation system assumes the user's preference over the item does not change over time. In many recommendation domains such as Youtube video recommendation or news recommendation, users constantly interact with the system with dynamic preference, and user feedback is instantly collected for improving recommendation performance. In these settings, it is essential for the recommendation method to adapt to the shifting preference patterns of the users. Other than the dynamic, the growing availability of data provides a unique opportunity for decision-makers to efficiently develop multi-dimensional (aka tensor) decisions for individuals. In this talk, I will discuss multi-armed bandit methods for online interactive recommendation with multi-dimensional actions.

Subspace-regularized Tensorial Parameter Estimation

Xin Zhang
Florida State University
E-mail: henry@stat.fsu.edu

Abstract: Statistical analysis of tensor data has many modern applications such as genetics, signal processing, neuroimaging, and recommender systems, and are also drawing increasing attention in high-dimensional statistics. In this talk, we propose a new framework for subspace regularized estimation of tensorial parameters in high dimensions. The proposed one-step refined estimator extends the recent algorithmic and theoretical developments in envelope methodology from vectors to tensors, and is applicable to various tensor regression and tensor discriminant analysis models.

High-dimensional Tensor Regression Analysis

Anru Zhang
University of Wisconsin-Madison
E-mail: anruzhang@stat.wisc.edu

Abstract: The past decade has seen a large body of work on high-dimensional tenors or multiway arrays that arise in numerous applications. In many of these settings, the tensor of interest is high-dimensional in that the ambient dimension is substantially larger than

the sample size. Oftentimes, however, the tensor comes with natural low-rank or sparsity structure. How to exploit such structure of tensors poses new statistical and computational challenges.

In this talk, we develop a novel procedure for low-rank tensor regression, namely Importance Sketching Low-rank Estimation for Tensors (ISLET) to address these challenges. The central idea behind ISLET is what we call importance sketching, carefully designed sketches based on both the responses and the structures of the parameter of interest. We show that our estimating method is sharply minimax optimal in terms of the mean-squared error under low-rank Tucker assumptions. In addition, if a tensor is low-rank with group sparsity, our procedure also achieves minimax optimality. Further, we show through numerical study that ISLET achieves comparable mean-squared error performance to existing state-of-the-art methods whilst having substantial storage and run-time advantages. In particular, our procedure performs reliable tensor estimation with tensors of dimension p = O(10^8) and is 1 or 2 orders of magnitude faster than baseline methods.

S031: Optimization Method and Theory for Big Data

Phase Transition of Landscape From Narrow to Wide Neural Networks

Ruoyu Sun
University of Illinois at Urbana-Champaign
E-mail: ruoyus@illinois.edu

Abstract: Wide neural networks are believed to have nice landscape, but what rigorous results can we prove, using just the condition of "wide"? We will show that: (i) From under-parameterized to over-parameterized networks, there is a phase transition from having sub-optimal basins to no sub-optimal basins. More specifically, for a dense set of activations and generic data, narrow networks have sub-optimal basins, while for all continuous activations, wide deep neural-nets have no sub-optimal basins. (ii) Over-parameterization alone cannot eliminate bad non-strict local minima, at least for a class of neurons. These results lead to a conjecture for lottery ticket hypothesis (Frankel and Carbin, ICLR'19): the lottery tickets in narrow networks are bad basins. These results are for un-regularized networks; time permitting, we will discuss results showing that with proper regularizers, even non-strict local minima can be eliminated.

Inference and Uncertainty Quantification for Noisy Matrix Completion

Yuxin Chen
Princeton University
E-mail: threshold.vincent@gmail.com

Abstract: Noisy matrix completion aims at estimating a low-rank matrix given only partial and corrupted entries. Despite substantial progress in designing efficient estimation algorithms, it remains largely unclear how to assess the uncertainty of the obtained estimates and how to perform statistical inference on the unknown matrix (e.g. constructing a valid and short confidence interval for an unseen entry).

This paper takes a step towards inference and uncertainty quantification for noisy matrix completion. We develop a simple procedure to compensate for the bias of the widely used convex and nonconvex estimators. The resulting de-biased estimators admit nearly precise non-asymptotic distributional characterizations, which in turn enable optimal construction of confidence intervals for, say, the missing entries and the low-rank factors. Our inferential procedures do not rely on sample splitting, thus avoiding unnecessary loss of data efficiency. As a byproduct, we obtain a sharp characterization of the estimation accuracy of our de-biased estimators, which, to the best of our knowledge, are the first tractable algorithms that provably achieve full statistical efficiency (including both the rates and the pre-constants). The analysis herein is built upon an intimate link between convex and nonconvex optimization.

On the Equivalence of Inexact Proximal ALM and ADMM for a Class of Convex Composite Programming

Xudong Li
Fudan University
E-mail: lixudong@fudan.edu.cn

Abstract: In this talk, we show that for a class of linearly constrained convex composite optimization problems, an (inexact) symmetric Gauss-Seidel based majorized multi-block proximal alternating direction method of multipliers (ADMM) is equivalent to an inexact proximal augmented Lagrangian method (ALM). This equivalence not only provides new perspectives for understanding some ADMM-type algorithms but also supplies meaningful guidelines on implementing them to achieve better computational efficiency.

S032: Recent Advances in Lifetime Data Analysis

Semiparametric regression analysis for serial gap times with competing events

Shu-Hui Chang
National Taiwan University
E-mail: shuhui@ntu.edu.tw

Abstract: Recurrent events are common in epidemiological and medical studies and usually followed by medical intervention or treatment in clinical practices. Serial gap times between consecutive interventions due to recurrence are natural outcomes of interest. The occurrence of recurrent events is often in conjunction with an event such as death which terminates the recurrent event process. After each intervention, the two types of subsequent event, recurrent and terminal events, compete with each other and investigating covariate effects on the risks of subsequent recurrent and terminal events is a primary focus in practice. Semiparametric regression models are introduced to model a sequence of episode-cause-specific hazards for serial gap times with episode-cause-specific covariate effects. We construct estimating equations for parameter estimation and study the asymptotic distributions of the proposed estimators without specifying the association pattern among serial gap times. Simulation studies are provided for examining the finite-sample properties of the proposed estimators. We apply the methods to papillary thyroid cancer data and investigate the covariate effects for different types and orders of events.

High dimensional data reduction in risk and survival data analysis

Catherine Huber
University Paris Descartes
E-mail: catherine.huber.carol@gmail.com

Abstract: Risk analysis is a topic of increasing importance in multiple fields like environment, technology and biomedicine.

In survival data analysis and reliability, one is interested in all risk factors that may accelerate or decelerate the life length of individuals or machines.

Now, as immense data bases (big data) are available, several types of methods are needed to deal with the resulting curse of dimensionality: on

one hand, methods that reduce the dimension while maximizing the infor-mation left in the reduced data, and then applying classical statistical models; on the other hand algorithms that apply directly to big data, i.e. artificial intelligence (machine learning), at the cost of a difficulty of interpretation in terms of the risk factors. Actually, those algorithms have a probabilistic interpretation. However, being often very good performers for prediction purposes, they lack explanatory interpretation.

We present here several methods for reducing the dimensionality of the data while maximizing the information relevant for the objective of the study, still present in the reduced data.

Statistical Analysis of Event Duration with Missing Origin

Yi Xiong
Simon Fraser University
E-mail: yi_xiong@sfu.ca

Abstract: Understanding the distribution of an event duration time is essential in many studies. The exact time to the event is often unavailable, and thus so is the full event duration. By linking information on longitudinal measures to the event duration via first-hitting-time model, we propose an estimation procedure for duration distribution and conduct semi-parametric regression analysis of event duration with missing origin. We establish the consistency and weak convergence of the proposed estimator and present its variance estimation. A collection of wildfire records is used to motivate and illustrate the proposed approach. The finite-sample performance of the proposed estimator is examined via simulation. Viewing the available data as interval-censored times, we show that proposed nonparametric estimator is valid and more efficient than the well-established Turnbull estimator, an alternative for these situations.

Variable screening with multiple studies and its application in survival analysis

Tianzhou Ma
University of Maryland School of Public Health
E-mail:tma0929@umd.edu

Abstract: Advancement in technology has generated abundant high-dimensional data that allows integration of multiple relevant studies. Due to huge computational advantage, variable screening methods based on marginal correlation have become promising alternatives to the popular regularization methods for variable selection. However, all screening methods are limited to single study so far. We consider a general framework for variable screening with multiple related studies, and further propose a novel two-step screening procedure using a self-normalized estimator for high-dimensional regression analysis in this framework. Compared to the one-step and rank-based procedures, our procedure greatly reduces false negative errors while keeping a low false positive rate. Theoretically, we show that our procedure possesses the sure screening property with weaker assumptions on signal strengths and allows the number of features to grow at an exponential rate of the sample size. Simulations and a real transcriptomic application illustrate the advantage of our method. Other than a linear model setting, our proposed framework is readily extensible to Cox model or threshold regression model in survival analysis for high-dimensional variable selection.

S033: Handling Complex Featured Data: Methods and Applications

Support Vector Machine with Graphical Network Structures in Features

Wenqing He
University of Western Ontario
E-mail: whe@stats.uwo.ca

Abstract: Machine learning techniques, regardless of being supervised or unsupervised, have attracted extensive research attention in handling data classification. Typically, among supervised machine learning algorithms, Support Vector Machine (SVM) and its extensions have been widely used in various areas due to their great prediction capability. These learning algorithms basically treat features of the instances independently when using them to do classification. However, in applications, features are commonly correlated with complex network structures. Ignoring such a characteristic and naively implementing the SVM algorithm may yield erroneous classification results. To address the limitation of the SVM algorithm, we propose new learning algorithms which accommodate network structures in the features of the instances. Our algorithms capitalize on graphical model theory and make use of the available R software package for SVM. The implementation of the proposed learning algorithms is computationally straightforward. We apply the new algorithms to analyze the data arising from a gene expression study.

Analysis of multivariate longitudinal data from eyes - microperimetry macular sensitivity loss in patients with Stargardt Disease

Xiangrong Kong
Johns Hopkins University
E-mail: xkong4@jhu.edu

Abstract: Microperimetry (MP) is a visual field test for measuring macular sensitivity of human eye. A MP test often involves testing of multiple locations in the visual field (e.g. 68 testing points), and the macular sensitivity is estimated as the mean sensitivity of the testing points. Macular sensitivity may be used as an endpoint in clinical trials for eye diseases. For Stargardt disease (the most common form of inherited juvenile macular degeneration), however, prior data have shown that the mean sensitivity (MS) change within one or two years was small, and thus MS may not be a sensitive measure to use as an endpoint for Stargardt trials. Clinically it has been hypothesized that the testing locations where macular lesions have already developed and are likely to expand, are the locations where function is mostly like to decline. Therefore, we are interested in using the point-level sensitivity data from MP test to characterize the points that are mostly likely to lose sensitivity and to estimate the sensitivity change in these points. Statistically, the data structure can be described as bivariate longitudinal multivariate outcomes, involving repeatedly annual MP tests for the 68 test points in both eyes. Such data structure entails complicated correlations, and simple analysis using generalized estimating equations or random effects models will not address the clinical hypothesis. We will use a point's neighboring points' sensitivity to characterize whether the point is at the lesion's edge, and use a hybrid modeling strategy based on Markov transition models together with pairwise composite likelihood for inference. The method is applied to the international multi-center Progression of Atrophy Secondary to Stargardt Disease (ProgStar) study to test the aforementioned clinical hypothesis and to determine whether the point-specific sensitivity in those points that had faster functional loss could be used as an endpoint for future Stargardt trials.

Dynamic risk prediction of a clinical event with sparse and irregularly measured longitudinal biomarkers

Yayuan Zhu
University of Western Ontario
E-mail: yayuan.zhu@uwo.ca

Abstract: Dynamic prediction of the risk of a clinical event by using longitudinally measured biomarkers or other prognostic information is important in clinical practice. It helps investigators understand the mechanism of disease progression and facilitates early prevention, decision making and resource planning. In this presentation, I will introduce the background and recent methodological development for dynamic risk prediction. We also propose a new class of semi-parametric landmark survival models applied to the context of sparse and irregularly measured longitudinal predictors. The model takes the form of linear transformation and allows all the model parameters to vary with the landmark time. This model includes many published landmark prediction models as special cases and imposes weaker assumptions. We develop a unified local polynomial estimation framework to estimate the unknown model components. We apply the proposed method to a data set from the African American Study of Kidney Disease and Hypertension (AASK) and predict individual patients' risk of end-stage renal disease (ESRD) or death as an illustration.

Degradation in common dynamic environments

Zhisheng Ye
National University of Singapore
E-mail: yez@nus.edu.sg

Abstract: Degradation studies are often used to assess reliability of products subject to degradation-induced soft failures. Because of limited test resources, several test subjects may have to share a test rig and have their degradation measured by the same operator. The common environments experienced by subjects in the same group introduce significant inter-individual correlations in their degradation, which is known as the block effect. In the present paper, the Wiener process is used to model product degradation, and the group-specific random environments are captured using a stochastic time scale. Both semiparametric and parametric estimation procedures are developed for the model. Maximum likelihood estimations of the model parameters for both the semiparametric and parametric models are obtained using an inexact block coordinate descent algorithm. Performance of the maximum likelihood estimators is validated through large sample asymptotics and small sample simulations. The proposed models are illustrated by an application to lumen maintenance data of blue light-emitting diodes.

S034: False discovery rate methodology

A Structure-Adaptive Learning Algorithm for Online False Discovery Rate Control

Wenguang Sun
University of Southern California
E-mail: wenguans@marshall.usc.edu

Abstract: Consider the online testing of a stream of hypotheses where a real-time decision must be made before the next data point arrives. The error rate is required to be controlled at {all} decision points. Conventional simultaneous testing rules are no longer applicable due to the more stringent error constraints and absence of future data. Moreover, the online decision-making process may come to a halt when the total error budget, or alpha--wealth, is exhausted. This work develops a new class of structure adaptive rules for online false discover rate control. The proposed algorithm is a novel alpha--investment rule that precisely characterizes the tradeoffs between different actions in online decision making. It captures useful structural information of the dynamic model, learns the optimal threshold adaptively in an ongoing manner and optimizes the alpha-wealth allocation in the next period. We present theory and numerical results to show that the proposed method controls the FDR at all decision points and achieves substantial power gain over existing online FDR procedures.

Controlling FDR while highlighting selected discoveries

Marina Bogomolov
Technion - Israel Institute of Technology
E-mail: marinabo@technion.ac.il

Abstract: "Modern scientific investigations often start by testing a large number of hypotheses by a False Discovery Rate controlling procedure, in order to identify the hypotheses that are promising for follow-up. In many cases, the set of discoveries is somewhat redundant, and it is subject to a second round of selection, where researchers identify the discoveries that better represent distinct findings for reporting and follow-up. For example, in genetic studies, if several genetic variants in a certain locus are identified as associated with the phenotype of interest, typically only the ""lead"" variant is reported, representing the entire locus. The guarantees of the FDR control for the initial set do not translate to this subset of reported discoveries. We show that if the rule defining how the discoveries will be filtered can be specified in advance, the Benjamini-Hochberg procedure can be modified to result in a focused set of discoveries with FDR guarantees. The proposed method allows researchers to curate rejections not only by subsetting, but also by prioritizing. We illustrate our methodology on a phenome-wide association study, where the hypotheses are structured as a tree.

Joint work with Eugene Katsevich and Chiara Sabatti"

Simultaneous confidence intervals in sequential estimation

Aaditya Ramdas
Carnegie Mellon University
E-mail: aaditya.ramdas@gmail.com

Abstract: A confidence sequence is a sequence of confidence intervals that is uniformly valid over all time. The advantage of a confidence sequence over pointwise intervals, is that the former are valid at stopping times (as well as post-hoc) but the latter are not. We show that at a price of about two (doubling of width), pointwise asymptotic confidence intervals can be extended to uniform nonparametric confidence sequences. Constructing the former at every time step guarantees FCR control, while constructing the latter each time step guarantees post-hoc FWER control.

Practical aspects of using False Discovery Rate

Yoav Benjamini
Tel Aviv University
E-mail: ybenja@gmail.com

Abstract: I shall propose guidelines to the reporting of results using of False Discovery Rate controlling methods, in scientific fields that retain high level of public interest, such as medicine and epidemiology. In particular I'll respond to the new statistical guidelines that were issued by the New England Journal of Medicine, and offer alternative guidelines.

S035: Multiple comparisons theory and applications

Closed testing and admissibility of procedures controlling false discovery proportions

Jelle Goeman

Leiden University Medical Center
E-mail: j.j.goeman@lumc.nl

Abstract: We consider a very general class of procedures controlling the tail probability of the number or proportion of false discoveries, either in a single (random) set or in several such sets simultanously. This class includes, among others, (generalized) familywise error, false discovery exceedance, simultaneous false discovery proportion control, and several selective inference methods. We put these procedures in a general framework, formulating all of them as giving a simultaneous lower confidence bound on the number of correctly identified discoveries in all possible subsets of the multiple testing problem. For such true discovery guarantee procedures we formulate both necessary and sufficient conditions for admissibility. First, we show that all such procedures are either a special case of closed testing, or can be uniformly improved by a closed testing procedure. Second, we show that a closed testing procedure is admissible as a true discovery guarantee procedure if and only if all its local tests are admissible. The practical value of our results is illustrated by giving a uniform (and substantial) improvement of a recently proposed selective inference procedure, achieved by formulating this procedure as a closed testing procedure.

All-Resolutions Inference for Brain Imaging

Aldo Solari
University of Milano-Bicocca
E-mail: aldo.solari@unimib.it

Abstract: Modern data analysis can be highly exploratory. Rather than a single well-defined research question there are tens of thousands of micro-questions, leading to tens of thousands hypothesis tests and tens of thousands of micro-inferences. These can be aggregated to larger-scale inferences in countless ways.

In brain imaging, for example, the brain is partitioned into tens of thousands of voxels, each of which may show activity as a response to stimulus. However, the unit of a voxel is arbitrarily determined by the measurement technique and does not represent a primary neural entity. The real research questions relate to patterns of activity at larger scales of aggregation, i.e. brain regions.

Researchers often highlight the patterns of brain activation suggested by the data, but false discoveries are likely to intrude into this selection. It is well-known that humans are very good at finding seemingly convincing patterns even in pure noise. How confident can the researcher be about a pattern that has been found, if that pattern has been selected from so many potential patterns?

We propose a novel approach - termed 'All Resolutions Inference' (ARI) - that delivers strong FWER control in any selected set of voxels. ARI allows a truly interactive approach to selective inference, that does not set any limits on the way the researcher chooses to perform the selection. The selection process does not have to be declared beforehand; it may be data-driven or knowledge-driven, or any mix of the two. Regardless of the selection process used, the researcher obtains a valid confidence bound for the proportion of truly active voxels in the final selected region.

Extrapolating expected accuracy for large multi-class problems

Yuval Benjamini
Hebrew University of Jerusalem ISRAEL
E-mail: yuvalbenj@gmail.com

Abstract: The difficulty of multi-class classification generally increases with the number of classes. This raises a natural question: Using data from a subset of the classes, can we predict how well a classifier will scale as the number of classes increases? In other words, how should we extrapolate the accuracy for small pilot studies to larger problems ?

In this talk, I will present a framework that allows us to analyze this question. Assuming classes are sampled from a population (and some assumptions about the classifiers), we can identify how expected classification accuracy depends on the number of classes (k) via a specific cumulative distribution function. I will present a non-parametric method for estimating this function, which allows extrapolation to K>k. I will show relations with the ROC curve. Finally, I hope to discuss why the extrapolation problem may be important for neuroscientists, who are increasingly using mutliclass extrapolation accuracy as a proxy for richness of representation.

This is joint work with Charles Zheng and Rakesh Achanta

S036: Advanced Topics in Survival Analysis

Left without being seen: The disappearance of impatient patients, combining current-status, right-censored and left-censored data

Yair Goldberg
Technion - Israel Institute of Technology
E-mail: yairgo@technion.ac.il

Abstract: I will present a survival-data setting that combines right-censored, left-censored, and current status data. The motivation is the challenge to estimate patience time, namely, the time till leaving without being served, of patients who arrive at an emergency department and wait for treatment. Three categories of patients are observed. The first category consists of patients who get service and thus their patience time is right-censored by the waiting time. The second category comprises those who leave the system and announce it, and thus their patience time is observed while the waiting time is right-censored. The third category consists of patients who leave the system without announcing it; their absence is hence revealed only when they are called to service, which is after they have already left; formally, their patience time is left-censored. The goal is to estimate the (im)patience distribution, based on these three data categories. I will present a novel methodology for distribution estimation using both parametric and nonparametric techniques. I will also present the performance of these estimators via simulated data and data from emergency departments.

Ventilation Prediction for ICU Patients with LSTM-based Deep Relative Risk Model

Bin Liu
Southwestern University of Finance and Economics
E-mail: liubin@swufe.edu.cn

Abstract: After admitted by the intensive care unit (ICU), a patient may experience mechanical ventilation (MV) if he/she suffers from acute respiratory failure. Vital signs and lab tests associated with the patient are typically recorded in a series over time.

We propose an LSTM-based deep relative risk model to quantify patients' time to occurrence of MV. The internal time-varying covariates motivate us to learn the ratio function via an LSTM net. The number of LSTM cells equals to the width of the sampling window; that is, the i-th cell of the LSTM net takes the patient's covariates of the time interval i as an input. A subsequent linear layer is used to summarize the hidden layers as the final partial likelihood contribution of each individual. Such an

architecture solves the survival analysis problem with internal time-dependent covariates in a nonparametric way. Our experiments based on the MIMIC-III database demonstrate it is a very promising approach to predicting the occurrence of MV.

Analysis of semi-competing risks data via bivariate longitudinal models

Daniel Nevo
Tel Aviv University
E-mail: danielnevo@gmail.com

Abstract: An example of semi-competing risk data analysis is the evaluation of risk factors for Alzheimer's disease and death, before and after the onset of Alzheimer's. Most existing methods treat the dependence between Alzheimer's and death as nuisance and restrict it to follow simple mathematical models. However, these methods may suffer from model misspecification of the dependence structure. Furthermore, information about the dependence, including its form, trajectory over time and how it depends on covariates can provide new clinical knowledge. Therefore, we propose a novel framework for analyzing semi-competing risks data by the means of bivariate longitudinal modeling. Our methods differentiate between local and global dependence. Local dependence captures the co-occurrence of Alzheimer's and deaths within a short period of time, while global dependence is the long-term effect of Alzheimer's on the risk of death. We incorporate flexible splines into our models to account for changes over time and develop a penalized maximum likelihood estimators and associated inference for the parameters of interest. Our methods are illustrated using the Adult Changes in Thought study.

Marginalized frailty-based illness-death model with application to biobank data

Malka Gorfine
Tel Aviv University
E-mail: malkago12@gmail.com

Abstract: The UK Biobank is a large-scale health resource comprising genetic, environmental and medical information on approximately 500,000 volunteer participants in the UK, recruited at ages 40--69 during the years 2006--2010. The project monitors the health and well-being of its participants. This work demonstrates how these data can be used to estimate in a semi-parametric fashion the effects of genetic and environmental risk factors on the hazard functions of various diseases, such as colorectal cancer. An illness-death model is adopted, which inherently is a semi-competing risks model, since death can censor the disease, but not vice versa. Using a shared-frailty approach to account for the dependence between time to disease diagnosis and time to death, we provide a new illness-death model that assumes Cox models for the marginal hazard functions. The recruitment procedure used in this study introduces delayed entry to the data. An additional challenge arising from the recruitment procedure is that information coming from both prevalent and incident cases must be aggregated. Lastly, we do not observe any deaths prior to the minimal recruitment age, 40. In this work we provide an estimation procedure for our new illness-death model that overcomes all the above challenges.

S037: New Analytical Solutions for Single-Cell and Functional Genomic Data

A statistical simulator scDesign for rational scRNA-seq experimental design

Jingyi Jessica Li
University of California, Los Angeles
E-mail: jli@stat.ucla.edu

Abstract: Motivation: Single-cell RNA sequencing (scRNA-seq) has revolutionized biological sciences by revealing genome-wide gene expression levels within individual cells. However, a critical challenge faced by researchers is how to optimize the choices of sequencing platforms, sequencing depths and cell numbers in designing scRNA-seq experiments, so as to balance the exploration of the depth and breadth of transcriptome information.

Results: Here we present a flexible and robust simulator, scDesign, the first statistical framework for researchers to quantitatively assess practical scRNA-seq experimental design in the context of differential gene expression analysis. In addition to experimental design, scDesign also assists computational method development by generating high-quality synthetic scRNA-seq datasets under customized experimental settings. In an evaluation based on 17 cell types and 6 different protocols, scDesign outperformed four state-of-the-art scRNA-seq simulation methods and led to rational experimental design. In addition, scDesign demonstrates reproducibility across biological replicates and independent studies. We also discuss the performance of multiple differential expression and dimension reduction methods based on the protocol-dependent scRNA-seq data generated by scDesign. scDesign is expected to be an effective bioinformatic tool that assists rational scRNA-seq experimental design and comparison of scRNA–seq computational methods based on specific research goals.

Availability and implementation: We have implemented our method in the R package scDesign, which is freely available at https://github.com/Vivianstats/scDesign.

Single-Cell Transcriptome and Regulome Data Integration

Weiqiang Zhou
Johns Hopkins Bloomberg School of Public Health
E-mail: wzhou14@jhu.edu

Abstract: New single-cell genomic technologies such as single-cell RNA-seq (scRNA-seq) and single-cell ATAC-seq (scATAC-seq) provide the capability for assaying the transcriptome (i.e., gene expression) and regulome (i.e., cis-regulatory element activities) of individual cells. To understand gene regulation in a biological system, one needs the information from both transcriptome and regulome. However, in most experiments, the same cell is only examined by either one of these technologies. Computational tools for integrating different types of single-cell genomic data are needed. Here, we present a new method that learns the connection between different data types based on public database and utilize such connection to integrate single-cell transcriptome and regulome data. We show that our method outperforms existing methods in aligning known cell types between scRNA-seq and scATAC-seq data. We further demonstrated our method by integrating the scRNA-seq data from Human Cell Atlas with public scATAC-seq data to study gene regulation in hematopoietic cell development.

SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing

Yuchao Jiang
University of North Carolina at Chapel Hill
E-mail: yuchaoj@email.unc.edu

Abstract: Whole genome single-cell DNA sequencing (scDNA-seq)

enables characterization of copy number profiles at the cellular level. This technology circumvents the averaging effects associated with bulk-tissue sequencing and increases resolution while decreasing ambiguity in tracking the evolutionary history of cancer. ScDNA-seq data is, however highly sparse and noisy due to the biases and artifacts that are introduced during the library preparation and sequencing procedure. Here, we propose SCOPE, a normalization and copy number estimation method for scDNA-seq data. The main features of SCOPE include: (i) a Poisson latent factor model for normalization, which borrows information across cells and regions to estimate bias, using negative control cells identified by cell-specific Gini coefficients; (ii) modeling of GC content bias using an expectation-maximization algorithm embedded in the normalization step, which accounts for the aberrant copy number changes that deviate from the null distributions; and (iii) a cross-sample segmentation procedure to identify breakpoints that are shared across cells from the same subclone. We evaluate SCOPE on a diverse set of scDNA-seq data in cancer genomics, using array-based calls of purified bulk samples as gold standards and whole-exome sequencing and single-cell RNA sequencing as orthogonal validations; we find that, compared to existing methods, SCOPE offers more accurate copy number estimates. Further, we demonstrate SCOPE on three recently released scDNA-seq datasets by 10X Genomics: we show that it can reliably recover 1% cancer cell spike-ins from a background of normal cells and that it successfully reconstructs cancer subclonal structure from ~10,000 breast cancer cells.

S038:Treatment Effects and Other Emerging Issues in Biomedical Data Science

Model-Free Causal Inference in Observational Studies

Ying Zhang
University of Nebraska Medical Center
E-mail: ying.zhang@unmc.edu

Abstract: Causal inference is a key component for comparative effectiveness research in observational studies. The inverse-propensity weighting (IPW) technique and augmented inverse-propensity weighting (AIPW) technique, which is known as a double-robust method, are the common methods for making causal inference in observational studies. However, these methods are known not stable, particularly when the models for propensity score and the study outcome are wrongly specified. In this work, we propose a model-free approach for causal inference. While possessing standard asymptotic properties, this method also enjoys excellent finite sample performance and robustness. Simulation studies were conducted to compare with the well-known IPW and AIPW methods for causal inference. A real-life example from an ongoing Juvenile Idiopathic Arthritis Study was applied for the illustration of the proposed method.

Estimating Treatment Effect under Additive Hazards Models with High-dimensional Covariates

Jue Hou
Harvard T.H. Chan School of Public Health
E-mail: dr.marquis.jue.hou@gmail.com

Abstract: Estimating causal effects for survival outcomes in the high-dimensional setting is an extremely important topic for many biomedical applications as well as areas of social sciences. We propose a new orthogonal score method for treatment effect estimation and inference that results in asymptotically valid confidence intervals assuming only good estimation properties of the hazard outcome model and the conditional probability of treatment. This guarantee allows us to provide valid inference for the conditional treatment effect under the high-dimensional additive hazards model under considerably more generality than existing approaches. In addition, we develop a new Hazards Difference (HDi) estimator. We showcase that our approach has double-robustness properties in high dimensions: with cross-fitting the HDi estimate is consistent under a wide variety of treatment assignment models; the HDi estimate is also consistent when the hazards model is misspecified and instead the true data generating mechanism follows a partially linear additive hazards model. We further develop a novel sparsity doubly robust result, where either the outcome or the treatment model can be a fully dense high-dimensional model. We apply our methods to study the treatment efect of radical prostatectomy versus conservative management for prostate cancer patients using the SEER-Medicare Linked Data.

stochastic search approach to identify subgroups with treatment benefit or harm

Changyu Shen
Harvard Medical School
E-mail: cshen1@bidmc.harvard.edu

Abstract: Existing statistical methods to identify sub-groups with differential treatment benefit/harm are either based on some parametric structure of the underlying data generation mechanism and/or are estimated through local optimization. We developed a nonparametric approach to identify subgroups through global optimization. Our approach is composed of two steps. In the first step, a discretization procedure creates a number of small sub-populations called “cells” with sufficient granularity, which serves as the building blocks of subgroup identification. In the second step, a simulated annealing algorithm is used to search for combinations of the cells that yield up to three groups: those deriving benefit from the treatment, those harmed by the treatment and the rest. Simulation studies are performed to evaluate the performance of this algorithm as compared with existing methods. A real data example is also presented.

Optimal Design and Analysis in Phase II Basket Like Trials

Fang Liu
Merck
E-mail: fang.liu11@merck.com

Abstract: The primary goal of an exploratory oncology clinical trial is to identify an effective drug for further development. To account for tumor indication selection error, multiple tumor indications are often selected for simultaneous testing in a basket trial. In this article, we propose optimal and minimax two-stage basket trial designs for exploratory clinical trials. Inactive tumor indications are pruned in stage 1 and the active tumor indications are pooled at end of stage 2 to assess overall effectiveness of the test drug. The proposed designs explicitly control the type I and type II error rates with closed-form sample size formula. They can be viewed as a natural extension of Simon’s optimal and minimax two-stage designs for single arm trials to multi-arm basket trials. A simulation study shows that the proposed design method has desirable operating characteristics as compared to other commonly used design methods for exploratory basket trials.

S039: Recent Advances in Statistical Methods for Single-cell Analysis

Statistical analysis of coupled single-cell RNA-seq and immune

profiling data

Hongkai Ji
Johns Hopkins Bloomberg School of Public Health
E-mail: hji@jhu.edu

Abstract: We present an analytical framework for analyzing coupled single-cell transcriptome (scRNA-seq) and T cell receptor sequencing (scTCR-seq) data. The framework provides key functions for preprocessing, aligning cells from different samples, detecting differential gene expression across biological conditions, analyzing sequence features in T cell repertoire, and linking sequence features to gene expression signatures. We demonstrate this framework by analyzing single-cell data both from public databases and from a neoadjuvant immunotherapy clinical trial for non-small cell lung cancer.

Gene expression imputation and clustering with batch effect removal in single-cell RNA-seq analysis by deep learning

Mingyao Li
University of Pennsylvania
E-mail: ahuwsj@126.com

Abstract: A primary challenge in single-cell RNA-seq (scRNA-seq) analysis is the ever increasing number of cells, which can be thousands to millions in large projects such as the Human Cell Atlas. Identifying cell populations becomes challenging in these data, as many existing scRNA-seq clustering methods cannot be scaled up to handle such large datasets. For large data, it is desirable to learn cluster-specific gene expression signatures from the data itself. Another challenge in large-scale scRNA-seq analysis is batch effect, which refers to systematic gene expression difference from one batch to another. Failure to remove batch effect can obscure downstream analysis and interpretation of results. In this talk, I will present a method for scRNA-seq analysis that enables gene expression imputation and clustering simultaneously through the use a deep learning algorithm. We further extend this method to incorporate known cell type information from a well-labeled source dataset through the use of transfer learning, a machine learning method that transfers knowledge gained from one problem to a different but related problem. Through comprehensive evaluations across many datasets generated in different tissues, species and protocols, we show that our methods can significantly improve clustering accuracy as compared to existing methods, and is capable of removing complex batch effects while maintaining true biological variations. We expect that, with the increasing growth of single-cell studies, our methods will offer a useful set of tools for clustering of these data.

SMNN: Batch Effect Correction for Single-cell RNA-seq data via Supervised Mutual Nearest Neighbor Detection

Yun Li
University of North Carolina
E-mail: yunli@med.unc.edu

Abstract: An ever-increasing deluge of single-cell RNA-sequencing (scRNA-seq) data has been generated, often involving different time points, laboratories or sequencing protocols. Batch effect correction has been recognized to be indispensable when integrating scRNA-seq data from multiple batches. A recent study proposed an effective correction method based on mutual nearest neighbors (MNN) across batches. However, the proposed MNN method is unsupervised in that it ignores cluster label information of single cells. Such cluster or cell type label information can further improve effectiveness of batch effect correction, particularly under realistic scenarios where true biological differences are not orthogonal to batch effect. Under this motivation, we propose SMNN which performs supervised mutual nearest neighbor detection for batch effect correction of scRNA-seq data. Our SMNN either takes cluster/cell-type label information as input, or, in the absence of such information, infers cell types by performing clustering of scRNA-seq data. It then detects mutual nearest neighbors within matched cell types and corrects batch effect accordingly. Our extensive evaluations in simulated and real datasets show that SMNN provides improved merging within the corresponding cell types across batches, leading to reduced differentiation across batches over MNN. Furthermore, SMNN retains more cell type-specific features after correction. Differentially expressed genes (DEGs) identified between cell types after SMNN correction are biologically more relevant, and the DEG true positive rates improve by up to 841%. SMNN is implemented in R, and freely available at https://yunliweb.its.unc.edu/SMNN/ and https://github.com/yycunc/SMNNcorrect.

S040: Survival Analysis and Beyond

Sparse Boosting for High-Dimensional Survival Data with Varying Coefficients

Jialiang Li
National University of Singapore
E-mail: stalj@nus.edu.sg

Abstract: Motivated by high-throughput profiling studies in biomedical research, variable selection methods have been a focus for biostatisticians. In this paper, we consider semiparametric varying-coefficient accelerated failure time models for right censored survival data with high-dimensional covariates. Instead of adopting the traditional regularization approaches, we offer a novel sparse boosting (SparseL2Boosting) algorithm to conduct model-based prediction and variable selection. One main advantage of this new method is that we do not need to perform the time-consuming selection of tuning parameters. Extensive simulations are conducted to examine the performance of our sparse boosting feature selection techniques.

Joint modeling of quality of life and survival data in palliative care studies

Zhigang Li
University of Florida
E-mail: zhigang.li@ufl.edu

Abstract: Palliative medicine is an interdisciplinary specialty focusing on improving quality of life (QOL) for patients with serious illness and their families. Palliative care programs are widely available or under development at US hospitals. In palliative care studies, often longitudinal QOL and survival data are highly correlated which, in the face of censoring, makes it challenging to properly analyze and interpret terminal QOL trend. Informative dropout in the study add another level of complication of the problem. To address these issues, we propose a novel statistical approach to jointly model the terminal trend of QOL and survival data accounting for informative dropout. We assess the model through simulation and application to establish a novel modeling approach that could be applied in future palliative care treatment research trials.

A spline-based nonparametric analysis for interval-censored bivariate survival data

Yuan Wu
Duke university

E-mail: yuan.wu@duke.edu

Abstract: In this manuscript we propose a spline-based sieve nonparametric maximum likelihood estimation method for joint distribution function with bivariate interval-censored data. We study the asymptotic behavior of the proposed estimator by proving the consistency and deriving the rate of convergence. Based on the sieve estimate of the joint distribution, we also develop an efficient nonparametric test for making inference about the dependence between two interval-censored event times and establish its asymptotic normality. We conduct simulation studies to examine the finite sample performance of the proposed methodology. Finally we apply the method to assess the association between two subtypes of mild cognitive impairment (MCI): amnestic MCI and non-amnestic MCI, for Huntington disease (HD) using data from a 12-year observational cohort study on premanifest HD individuals, PREDICT-HD.

Causal Effects on Birth Defects with Missing by Terathanasia

Andrew Ying
University of California
E-mail: aying9339@gmail.com

Abstract: We are interested in the causal effects of the etanercept on the major birth defects among women diseased when compared to women with the same disease but without exposure during pregnancy, up to some sub-populations at most. We use the data from the prospective observational pregnancy cohort studies where women with certain autoimmune diseases are exposed to etanercept during pregnancy. Like the usual observational studies, this data suffers from several complications. In particular, the outcomes are sometimes missing due to the spontaneous abortion (SAB), in that the aborted fetuses are often not examined for defects. The theory of terathanasia indicates that the aborted fetuses are more likely to be malformed than the live born infants, leading to the missing not at random. Additionally, there are left truncations. Limited by the moderate sample size, models that are capable of handle all these complications while being parsimonious are demanded. We adopt the Rubin causal model citep{holland1986statistics} to formulate the causal effects. For the average exposure effect, we adopt the inverse probability weighting (IPW) method together with the selection model where the time to spontaneous abortion is modeled using the survival techniques to handle the left truncation in the data. In addition, the effect of etanercept on the birth defects was often presented for the live born stratum in practice, itself a post-exposure variable which invalidated any causal interpretations. We explore the principal strata approach citep{frangakis2002principal}, given the very limited sample size of such studies. The population is stratified by whether an individual will experience the SAB if taking the medicine or not.

We adopt the likelihood-based analyses estimated by the expectation substitution (ES) algorithm citep{elashoff2004algorithm}. The inferential results are obtained by virtue of the analog Louis' formula. The treatment effect is positive on the whole population, except which is not significant. We find that the treatment has a positive effect on the subpopulation who will never experience SAB not matter what treatment they receive while showing a negative effect on those who will always experience SAB. We also conduct the sensitivity analyses to assess the robustness of our conclusion by varying the effect of ``terathanasia".

S041: Novel Statistical Approaches to Investigate Cancer Immunotherapy

A Transcriptome Based Nonparametric Method to Deconvolute Immune Cells and Cancer Subtypes

Guoshuai Cai
University of South Carolina
E-mail: caigs.whu@gmail.com

Abstract: The molecular characterization of immune cells as well as cancer subtypes is important for understanding the disease developments and searching effective treatment. A handful models and methods have been developed and shown power in estimating the composition of immune cells or cancer subtypes based on several signatures or the whole transcriptome. However, signature-based methods are easily influenced by between-dataset variations while transcriptome-based methods suffer from measurement noises. Weighted regression improves the estimation but require the distribution assumptions and a sufficient knowledge for weight inference. Given the high complexity of biological system and the small size of experiments, we proposed a non-parametric method by integrating the information from both signatures and transcriptome, which showed significantly improved robustness on simulation and real datasets.

A statistical framework to investigate molecular mechanisms associated with tumor microenvironment

DONGJUN CHUNG
Medical University of South Carolina
E-mail: chungd@musc.edu

Abstract: During the last decade, there have been tremendous achievements in cancer immunotherapy. Among those, immune checkpoint blockades, such as Anti-PD1, have completely changed the therapeutic approaches for many type of cancer. However, a significant heterogeneity in the efficacy of these immune checkpoint blockades has been reported and the molecular basis related to such differences have not been thoroughly investigated yet. Compositional data analysis recently received significant attention with the emergence of big compositional data such as microbiome and immune cell composition data. In this presentation, I will discuss our recent work on a Bayesian regression framework for the compositional data. This approach allows us to consider correlation among compositional outcomes, identify key covariates associated with the compositional outcomes, and utilize various prior biological knowledge. We applied the proposed statistical method to the immune-genomic data of the Immune Landscape of Cancer.

Phase I/II dose finding interval design for *Immunotherapy*

Yeonhee Park
Medical University of South Carolina
E-mail: parkye@musc.edu

Abstract: Immunotherapeutics have revolutionized the treatment of metastatic cancers and are expected to play an increasingly dominant role in the treatment of cancer patients. Recent advances in checkpoint inhibition show promising early results in a number of malignancies, and several treatments have been approved for use. However, the immunotherapeutic agents have revealed to have different toxicity profiles and mechanism of antitumor activity from the cytotoxic agents, and many limitations and challenges encountered in the traditional paradigm were recently pointed out for immunotherapy. Our methods address the difficulty to identify the relationship between immunotherapeutic exposure and clinical outcomes and determine optimal biological dose of immunotherapeutics by effectively utilizing toxicity, immune response, and tumor response. Moreover, we propose an algorithm to allocate the dose for next cohort

which makes dose transition safer and more appropriate by prioritizing the safety over efficacy outcomes, which is analogous to the rationale of phase-I-and-then-II. Simulation studies show that the proposed design has desirable operating characteristics compared to existing dose-finding designs. It also inherits strengths of interval designs to have superior performance with the simplicity of the algorithm based on multiple outcomes.

Sparse LDA with Network-Guided Block Covariance Matrix

Jin Hyun Nam
Medical University of South Carolina
E-mail: namj@musc.edu

Abstract: In the high-dimensional setting, linear discriminant analysis is faced with two challenges, namely singularity of the covariance matrix and difficulty of interpreting the resulting classifier. Although several methods have been proposed to address these problems, most of them did not take into account dependency between variables and efficacy of selected variables appropriately and they focused only on identifying a parsimonious set of variables maximizing classification accuracy. To address this limitation, here we propose a new approach that directly estimates the sparse discriminant vector without need of estimating the whole inverse covariance matrix, which can be formulated as a quadratic optimization problem. Furthermore, this approach allows to integrate external information to guide the structure of covariance matrix. We applied the proposed method to the transcriptomic study that aims to identify genomic markers predictive of the response to cancer immunotherapy, where the covariance matrix was constructed based on the communities identified from gene-gene networks.

S042: Recent Advancement in Biostatistics Methodology

Joint analysis of multiple longitudinal and survival data measured on nested time-scales: an application to predicting infertility

Rajeshwari Sundaram
National Institutes of Health
E-mail: sundaramr2@mail.nih.gov

Abstract: "Fertility-tracker apps are now widely used by couples attempting to get pregnant. Thus, providing a rich source of data that are large in magnitude compared to the more traditional scientifically designed pregnancy studies. Motivated by these rich data sources, we aim to build models for predictions of individual time-to-pregnancy based on joint models of underlying biology and behavior of couples. We will first discuss the joint modeling of longitudinal binary data (ie, intercourse pattern of couples), highly skewed longitudinal process (ie, menstrual cycle lengths, proxy for her reproductive health) and a discrete survival time (ie, time-to-pregnancy).

The intercourse observations are a long series of binary data with a periodic probability of success and the amount of available intercourse data is a function of both the menstrual cycle length and TTP. Moreover, these variables are dependent and observed on different, and nested, time scales (TTP measured in months, length of each menstrual cycle in months while intercourse measured on days within a menstrual cycle) further complicating its analysis. Here, we propose a semi-parametric shared parameter model for the joint modeling of the binary longitudinal data (intercourse behavior), skewed continuous longitudinal process (menstrual cycle) using a mixture distribution and the discrete survival outcome (TTP). Finally, we develop couple-based dynamic predictions to assess the risk for infertility. We will discuss computational methods used to make the model fitting fast as well as talk about how our approach can be used to model data collected from app-based fertility trackers."

Modeling and Correlation Estimation for Bivariate Recurrent Event Processes

Mei-Cheng Wang
Department of Biostatistics, Johns Hopkins University
E-mail: mcwang@jhu.edu

Abstract: Bivariate or multivariate recurrent event data are often collected in longitudinal studies as the primary outcome measurements for research. We consider modeling and correlation structure for bivariate recurrent events, where the association between two types of recurrent events is characterized by frailty processes and hence allows for time-dependent association. This forms a contrast with those conventional models for bivariate recurrent events where the association is characterized solely by a baseline frailty variable. Composite likelihood approaches are developed to estimate parameters in the joint rate models in semiparametric settings. The proposed models and methods can be used to identify biomarkers or risk factors for recurrent events that could be used to tailor preventive strategies and treatment plans. An analysis of stroke data is presented to illustrate the applicability of the proposed methods.

Analysis of competing risks data with dependent truncation

Yu Jen Cheng
National Tsing Hua University
E-mail: ycheng@stat.nthu.edu.tw

Abstract: In this work, two specific challenges are encountered in the analysis of competing risks data under prevalent sampling. First, because the observation of failure times is subject to left truncation, the sampling bias extends to the failure type which is associated with the failure time. An analytical challenge is to deal with such sampling bias. Second, the cumulative incidence function is allowed to have a temporal trend. Mixture model approaches are proposed to address these two analytical challenges on the basis of prevalent survival data. The proposed approaches are examined through simulation studies and illustrated by using a real data set.

Biomarker Guided Phase II Two-Stage Design for Targeted Therapy

Zheyu Wang
JHU
E-mail: wangzy@jhu.edu

Abstract: Successful development of targeted therapy often relies on appropriate sub population selection. Biomarker assessments are increasingly involved in such trials. Most of biomarker-guided designs assume that a biomarker cutoff value has been proposed. In practice however, biomarker development often lags behind therapeutic development and a cutoff value is often difficult to determine at trial planning stage. On the other hand, designs that allows for biomarker cutoff determination often consider a phase II/III setting where the sample size is larger. These designs also primarily aim to claim efficacy in the entire population. With the development of targeted therapy and the challenge in speedy development of associated biomarker, we expect more and more experimental drugs to be most beneficial only in subgroup of patients and this group is often unknown at the time of trial planning. A design that can properly select the subgroup and adequately power the test in this efficacy

subgroup is desirable. In this talk, we discuss a two-stage design based on Bayesian decision-theoretic approach for this purpose. We also discuss the sample size allocation between both stages in this design.

S043: SIBS Invited Session: Recent Advancement in Biostatistics Methodology

Hierarchical Bayesian Spatio-Temporal Models with Application to Birds Population Spread

Xuejing Meng
1. Simon Fraser University 2. Hubei University of Economics
E-mail: mengxuejing18@163.com

Abstract: Many researcher in environmetrics are interested in modeling evolution of certain variables, such as wind, temperature, moisture, population, and the associated inference methods. Observations on the variables are often with spatio-temporal features. Extending the diffusion-reaction partial differential equation (PDE) in the literature (e.g. Wikle 2003), we formulate the population bird spread using advection-diffusion-reaction partial differential equation (PDE) in a hierarchical Bayesian framework. The model can account the existence of possible trend and give the diffusion process of birds under this trend. We consider a Poisson response with the trend coefficients and spatially varying diffusion coefficients. Moreover, the increasing term is assumed to follow an advection-diffusion-reaction PDE. This mimics realistically the birds population spread process. We illustrate the approach via both simulation and a set of real data.

Inferring Longitudinal antiretroviral drugs effects on mental health in people with HIV

Yanxun Xu
Johns Hopkins University
E-mail: yanxun.xu@jhu.edu

Abstract: The effects of antiretroviral (ART) drugs for people living with HIV (PLWH) on mental health are inconsistent. Given the heterogeneous nature of both ART drugs and the presentation of depressive symptoms, newer approaches are necessary for guiding clinical practice. Since ART-related depression would be heterogeneous among HIV patients depending on their differences in numerous factors including demographics and clinical variables, we develop a new Bayesian semiparametric graphical model with nodes representing drugs and depression items, and weighted edges representing their relationships. The weights indicate the strength of the drug-depression relationships and can vary across different visits and different patients. The effective and reliable modeling and prediction will help elucidate the treatment-depression relationship and guide the clinicians in making more informed decision for patients.

Use of Multistate Model for Multiple Endpoints in Oncology Clinical Trials Analysis and Designs

Chen Hu
Johns Hopkins University
E-mail: chu22@jhmi.edu

Abstract: In oncology clinical trials, disease progression and mortality are typically captured through a series of sequentially observed events, such as cancer recurrence and deaths. The relationship between covariate (e.g., therapeutic intervention and prognostic markers), recurrence, and death is often of interest, as it may provide key insights of optimal treatment decisions and future study designs. Analysis of these multiple endpoints however can be complicated due to censoring, under-reporting of intermediate event and potential correlation between events. We focus on how multistate model framework and semiparametric regression model can be used to handle these challenges and provide insights on evaluating treatment effect, identifying potential prognostic and predictive covariates on disease progression and mortality, and facilitate late-phase clinical trial designs. Asymptotic results and numerical examples based on Monte Carlo simulations and real trial data are presented.

An ADMM Algorithm for Distributed Sparse Optimal Scoring Classification

Yuanshan Wu
Zhongnan University of Economics and Law
E-mail: shan@whu.edu.cn

Abstract: This talk discussses the classification analysis with huge sample size and high dimensionality. Due to the limited storage of a single machine, the whole dataset are usually stored across multiple ones. It makes the traditional classification methodologies unapplicable as well as introduces additional challenges such as computation complexity and communication cost. We propose a distributed sparse optimal scoring classification based on the alternating direction method of multipliers algorithm. Specifically, by imposing consensus-based constraint, the estimates in each machine are forced to be equal and the optimization problem can be separately performed on each machine, making the parallelized computation feasible. We show that at a linear rate the proposed estimates converge to the global ones which are obtained by assuming a single super-machine could store the full data. The merits of algorithms corroborating the global optimality and convergence are demonstrated through both simulated and real data sets.

S044: Data science and statistics in IT companies

Semantic clustering of YouTube videos

Ying Liu
Google
E-mail: yingliug@google.com

Abstract: Recent developments in image and video analytics allow tagging of elements that appear in the videos. However it could be difficult to identify the semantic meaning of videos solely based on such tags. In this talk I will discuss methods and results of semantic clustering of videos, which can be useful in follow-up analyses.

Comparison Studies of Multi-Armed Bandit Algorithms for Display Advertising Optimization

Wanghuan Chu
Google
E-mail: dqchuwh@gmail.com

Abstract: The classical multi-armed bandit problem has gained increasing popularity and attention in both academia and industry. The core challenge consists of a good balance between exploration and exploitation, with the intention to maximize the accumulated rewards over the long run. Online display advertising is one of the important application areas of multi-armed bandit. Specifically, given a specific group of ads (referred to as an ad group), we want to allocate total N impressions (as N pulls) to the K different ads (as K arms) in the particular group, with the attempt to maximize expected accumulated rewards (e.g. clicks). The main contributions of this project include: 1) we setup the simulation studies to characterize the special features for online advertisement optimization, providing guidance to choose methods and the way to tune their corresponding parameters for practitioners who encounter similar problems;

2) we study different factors affecting performance; 3) we also conduct our comparison experiments with non-stationary scenarios, as it's more similar to real applications.

Inferring Impact Direction Graphs from Large Scale Online User Engagement Data

Yuxiang Xie
Snap Inc.
E-mail: yxie@snap.com

Abstract: Nowadays many companies of online services or mobile apps are using user engagement based metrics as the pointers toward the North Star (i.e. the success) of the business. However, many commonly used North Star metrics, such as Daily Active Users (DAU), Monthly Active Users (MAU), long-term revenue per user, are often not useful for day-to-day decision making because either they are insensitive to small and incremental product improvements, or their short-term movements mis-align with the real user experience in the long-term [1]. Thus it has been critical for online service or mobile app developers to find user engagement metrics that are actionable enough in the short term, and at the same time, are constantly the impact drivers for the growth of the North Star metrics in the long term. In this work, we propose a fast and scalable method based on Concave penalized Coordinate Descent with reparametrization (CCDr) [2] to learn the skeleton of the impact relationships among various user engagement metrics of an online service or mobile app. By applying the method on more than 1,500 A/B tests data of Snapchat, we construct impact direction graphs that deliver clear and useful insights on the impact relationship among the user engagement metrics, and effectively identify the metrics that eventually drive users' long-term retention in spite of the complex user activities in using Snapchat.

[1] Deng, A. and Shi X. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. SIGKDD 2016.

[2] Aragam, B. and Zhou, Q. Concave penalized estimation of sparse Gaussian Bayesian networks. Journal of Machine Learning Research, 16:2273–2328, 2015

Statistics and Big data at Google

Lu Zhang
Google
E-mail: luzhang@google.com

Abstract: Provide an overview of Big data at Google, and typical statistical problems to solve, including experimentation, causal modeling for observation data, and surveys.

S045: Recent Development in Risk Measure and Its Application

Measuring systemic risk contagion effect of the banking industry in China: A directed network approach

Zisheng Ouyang
Hunan University Of Technology and Business
E-mail: ouyang_zs@163.com

Abstract: To capture the impact of investor sentiment on the risk contagion of financial institutions and the potential tail risks caused by the financial network structure, this paper uses the directed network approach to measure systemic risk contagion effect of the Chinese banking industry. Considering the nonlinear mechanism of financial risk mutation and contagion, we use the linear quantile lasso regression and local polynomial method to estimate the TENET model which is a new method of systemic risk measurement, and construct a weighted directed network. Moreover, we study the directed network from different perspectives, analyze the financial risk contagion effect and the influence of investor sentiment on the financial risk contagion, and identify systemically important financial institutions. Using 16 listed banks in China as samples, we find that: (1) With the spread of crisis, the entire financial system becomes more closely related, and the total network connectivity continues to rise until it reaches a maximum value. (2) The total network connectivity and the average value (systemic risk) have the same upward or downward trend, but the average value lags behind the total network connectivity. (3) The current bank has characteristics of "too big to fail" and "too contact to fail".

Keywords: Systemic risk; Contagion effect; Investor sentiment; Directed network approach; CoVaR model

The finite-time ruin probability of a discrete-time risk model with subexponential and dependent insurance and financial risks

Shijie Wang
Anhui University
E-mail: ahuwsj@126.com

Abstract: Consider a discrete-time risk model with insurance and financial risks in a stochastic economic environment. Assume that the insurance and financial risks form a sequence of independent and identically distributed random vectors with a generic random vector following a wide type of dependence structure. An asymptotic formula for the finite-time ruin probability with subexponential insurance risks is derived. In doing so, the subexponentiality of the product of two dependent random variables is investigated simultaneously.

Robust portfolio with multi-objective optimization model under high-dimensional scenarios

Xia Zhao
Shanghai University of International Business and Economics
E-mail: zhaoxia-w@163.com

Abstract: This paper aims to study robust portfolio with mean-variance-CVaR criteria for high-dimensional data. Combining different estimators of covariance matrix, computational methods of CVaR and regularization methods, we construct five progressive optimization problems with short-selling allowed. The impacts of different methods on out-of-sample performance of portfolios are compared. Results show that the optimization model with well-conditioned and sparse covariance estimator, quantile regression computational method for CVaR and weighted LASSO performs best, which servers for stabilizing the solution and also encourages a sparse portfolio

S046: Recent method and technique developments in genomics and drug safety

Group-level network inference via l_0 shrinkage and graph combinatorics

Shuo Chen
University of Maryland, School of Medicine
E-mail: chenshuochen@gmail.com

Abstract: We consider group-level statistical inference for networks, where the outcome variables of each subject are multivariate edges in an adjacency matrix. We assume the nodes of adjacency matrices are identical across all subjects and the goal is to identify and statistically test whether edges in some subnetworks that are associated with the covariates of

interest. We propose a group level network inference (GLEN) framework to extract the subgraphs where edges are likely to be related to the covariate via l_0 norm regularization and perform statistical tests on the detected subgraphs by graph combinatorics. Theoretical properties of the novel objective function and network-level inference are provided. We apply the proposed method to a brain connectomic study to identify the subnetworks of brain-connectome that are associated with brain diseases. In addition, we perform extensive simulation studies. The results demonstrate the proposed method outperform existing multivariate statistical methods by simultaneously improve false positive and false negative discovery rates and significantly increase replicability.

Integrated sequencing analysis for virus detection in Human disease

Lijun Zhang
Pennsylvania State University College of Medicine
E-mail: lzhang6@pennstatehealth.psu.edu

Abstract: Virus infection and its interaction with the host genome in human disease remains challenges in term of structural variation in the human genome. The limitation is that the major tool for interrogating genome organization – Next Generation Sequencing (NGS) is inadequate for resolving large structural change in the genome or the genome assembly. Here, we adopted a novel technique, the optical genomic mapping (OGM) in conjunction with NGS to detect virus and to check the molecular structure of virus associated with human genome. Additionally, we investigate the instability correlated with integrated and episomal virus genomes in cancer cell. Our objective is to analyze the state of the viral genome in the cancer cell, analyze the effect of viral genome integration on adjacent host genomic structure and correlate findings to patient clinical outcomes.

Our results show that established OGM and NGS can identify the nature of viral association with the genome in cancer cells and in the case of integration, determine the structure of viral integration sites. Our preliminary data from human sample suggests viral integration may be associated with greater levels of genomic instability in comparison to episomal viral structures.

Statistical test of structured continuous trees based on discordance matrix

Lin Wan
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
E-mail: lwan@amss.ac.cn

Abstract: Cell fate determination is a continuous process in which one cell type diversifies to other cell types following a hierarchical path. Advancements in single-cell technologies provide the opportunity to reveal the continuum of cell progression which forms a structured continuous tree. Computational algorithms, which are usually based on a priori assumptions on the hidden structures, have previously been proposed as a means of recovering pseudo-trajectory along cell differentiation process. However, there still lack of statistical framework on the assessments of intrinsic structure embedded in high-dimensional gene expression profile. We propose an adaptive statistical framework, termed structured continuous tree (SCTree), to test the intrinsic structure of a high-dimensional single-cell dataset. SCTree test is developed on the basis of the tools governing metric geometry and random matrix theory. We show that the SCTree test is most powerful when the signal-to-noise ratio exceeds a moderate value. We also demonstrate that SCTree is able to robustly detect linear, single and multiple branching events with simulated datasets and real scRNA-seq datasets.

Bayesian Modeling of Rare Events Data with Missing Not At Random

Shouhao Zhou
The Pennsylvania State University
E-mail: szhou1@phs.psu.edu

Abstract: This study is motivated by a meta-analysis for drug safety in clinical trials, when a large number of rare adverse events (AEs) are not reported if they are less frequently observed. As a typical missing not at random problem, is the censored information ignored, the inference on incidence rate of AEs would be overestimated by nearly 40%. We propose a modified Bayesian multilevel logistic regression model to accommodate the censored sparse binomial event data, and implement in JAGS based on a tailored modeling strategy. We conduct simulation studies to examine the performance of our proposed Bayesian model compared to other popular methods in finite samples under four scenarios. The proposed approach is illustrated using data from a recent meta-analysis of 125 clinical trials involving PD-1/PD-L1 inhibitors with respect to their toxicity profiles.

S047: Recent Advances in Analytic Methods for Sequencing and Biobank Data

Bayesian Covariate-dependent Gaussian Graphical Model

Yingying Wei
The Chinese University of Hong Kong
E-mail: yweicuhk@gmail.com

Abstract: There has been active research on estimating a single Gaussian graphical model for a set of samples. However, when there exists heterogeneity among the samples, learning a single graphical model for all of the samples can lead to many spurious edges. The recent emerging methods on joint modeling of multiple graphical models allow graphical structures to change with univariate, categorical covariates. However, they cannot handle continuous covariates, let alone multiple covariates. The early proposed frequentist approach to linking multiple covariates to graphical structures must first partition the space of covariates and then separately learn graphs on each portion of the data, resulting in unstable estimators and a loss of interpretability for the covariates.

Here, we propose a novel Bayesian framework to study how the graphical structures change with covariates for Gaussian graphical models. Our proposed method can handle all types of covariates, borrow strength across the whole covariate space to improve edge detection, and provide direct interpretation for effects of covariates on each edge of the graph. We develop an efficient parallel Markov chain Monte Carlo algorithm to conduct posterior inference. We applied the proposed method to study how gene regulatory networks vary across different types of covariates such as age and disease categories.

Statistical assessment of depth normalization methods for microRNA sequencing

Jian Zou
University of Pittsburgh
E-mail: jian.zou@pitt.edu

Abstract: Quality data is the foundational cornerstone for reliable scientific findings in evidence-based medical research. It is widely accepted that a

crucial step to derive high-quality genomics data is to identify data artifacts caused by systematic differences in the processing of specimens and to remove these artifacts by data normalization. One major and unique aspect of RNA sequencing data normalization is the 'depth of coverage'. Statistical methods for depth normalization have been recently developed, including both simple rescaling-based methods and regression-based methods. Many of these normalization methods rely on the presupposition that variations in the assumed scaling factor or in the projection of the assumed regression function are solely due to data artifacts and should be removed. MicroRNAs are a unique class of small RNAs regulating gene expression and closely linked to carcinogenesis. They are low-complexity molecules (that is, a small number of molecules expressed dominantly) that tend to be expressed in a tissue-specific manner, especially in heterogeneous samples such as tumors. As a result, the assumption of depth normalization methods may not hold for microRNA sequencing. We performed a study to assess the performance of existing depth normalization methods on identifying disease-relevant microRNAs using both a pair of datasets on the same set of tumor samples and data simulated from the paired datasets under various scenarios of differential expression. In this talk, we will report our findings from this study.

Estimating the effect of covariates on the correlation between bivariate failure times.

Sean Devlin
Memorial Sloan Kettering Cancer Center
E-mail: sedevlin@gmail.com

Abstract: The correlation between two time-to-event outcomes, such as the composite endpoint of event-free survival and the terminal endpoint of overall survival, can be estimated as an overall measure for a study population. However, this correlation may vary based on the underlying clinical and genomic characteristics of the patient. In this talk, we propose a nonparametric regression approach to estimating and visualizing the correlation of the bivariate failure times across the covariate space. The operating characteristics of this approach are evaluated using a simulation study, and the methodology is illustrated in two cancer data sets.

Mediation Analyses of Ultraviolet, Air Pollution, and Structural Variations using the Taiwan Biobank

En Yu Lai
Institute of Statistical Science, Academia Sinica
E-mail: junelai@webmail.stat.sinica.edu.tw

Abstract: Structural variation is a DNA region that shows changes in copy number, sequence orientation or chromosomal location. Previous studies have suggested a link between air pollution and genetic variation in animal experiments and longitudinal studies, but the sample size is rather limited. It is imperative that a population-based study is conducted to document the potential hazard of environmental exposures such as air pollution and ultraviolet to the human genome and health. The Taiwan Biobank has been collecting biological specimens and conducts the whole-genome sequencing in order to build the reference genome of the Taiwanese population. In this study, we aim to characterize the causal relationship between ultraviolet, air pollution and structural variations. We applied a mediation model to describe the influence of ultraviolet toward structural variants through air pollution. The preliminary results showed a strong effect from ultraviolet to structural variants mediated by air pollution. Validation studies are needed to confirm this interesting finding.

S048: Statistics decision in Drug Development

A decision-theoretic framework for multiple testing controlling the familywise expected loss

Xiaolei Xun
Fudan University
E-mail: xiaolei_xun@fudan.edu.cn

Abstract: We consider the problem of testing multiple null hypotheses where a decision to reject or retain is to be made for each individual hypothesis. Based on the decision-theoretic framework, we propose to control the familywise expected loss instead of the conventional familywise error rate (FWER). Various loss functions can be adopted and the FWER is seen to result as a particular choice of the loss function. We search for decision rules that satisfy certain optimality criteria within a broad class of rules for which the expected loss is bounded by a pre-specified threshold under any parameter configuration. This approach is different from the canonical decision theory of maximizing a single utility function, but in analogy to classical hypothesis testing. We illustrate the methods with the problem of establishing efficacy of a new medicinal treatment in non-overlapping subgroups of patients.

Sample Size Determination Concerning Decision Making in Clinical Trials - Two Case Studies

Julie Ma
Gilead Sciences, Inc
E-mail: julieggma@gmail.com

Abstract: A pivotal aspect of planning a clinical study is the calculation of the sample size. The calculation of an adequate sample size is crucial, a process by which we calculate the optimum number of participants required to be able to arrive at ethically and scientifically valid results. Generally, the sample size is a function of significance level, power, expected effect size, drop-out rate, allocation ratio, and the objective and design of the study. In reality, the sample size needs to take into account of the decisions to be made during the course of the study and at the end of the study. In many cases, the objective of the drug development is not only to exceed placebo, but also to identify a competitive drug candidate, especially for a disease area that the market is very crowded. Therefore, the sample size will not be simply based on significant p values, but be determined by quantitative decision criteria comparing to competing drugs. During the course of the study, interim analyses for futility are often performed. The sample size should be planned sufficiently so that the interim decision making is sound given various assumptions. At the end of phase 2, a decision to initiate phase 3 studies is usually made based on the predicted probability of success of which the size of the phase 2 trial is a key component. In this talk, two real examples will be given to illustrate these ideas for the sample size determination concerning decision making.

Gating criteria using Bayesian approach in early phase study

Wenxin Liu
Roche China
E-mail: wenxin.liu@roche.com

Abstract: This presentation will introduce the gating criteria using Bayesian approach in early phase study.

Usually gating criteria is prespecified before readout of a small sample size of patients' data. Based on the prespecified gating criteria, the sponsor can develop decision making strategy, including Pivotal Go, Go, No Go. In this presentation, a binary endpoint ORR is considered.

Quantitative decision making in preclinical drug discovery.

Xikun Wu
BeiGene
E-mail: xikun.wu@beigene.com

Abstract: Preclinical discovery is an important component in the drug development life cycle. From target identification and validation, high throughout screening, lead optimization to clinical candidate selection, statistician need to integrate in the project team to understand the subject matter thus to utilize the statistics expertise and translational thinking to support quantitative decision making in each step. This talk will give an overview of the preclinical drug discovery. Critical question in each stage will be introduced with some examples of statistics experiment designs and data analysis.

S049: Utilization of Big Data in Precision Medicine

Subgroup Discovery Using Consensus Clustering Methods

J. Richard Landis
University of Pennsylvania
E-mail: jrlandis@pennmedicine.upenn.edu

Abstract: Discovery of patient subtypes with differential profiles of risk for multiple outcomes is essential for improving health promotion and precision medicine strategies. A K-means clustering algorithm, using an average linkage method, was applied to I item clusters as variables, and was repeated within 1,000 randomly selected subsamples of size 80%N, utilizing R Bioconductor ConsensusClusterPlus (Wilkerson, Hayes and Neil (2010)). An evaluation criterion based on the "Proportion of Ambiguous Pairs within Clusters" (PAC) was used to generate a mean consensus score for each of the K clusters. A new sequential hypothesis testing approach was utilized to determine the ideal number of clusters. These methods were applied to deep phenotyping data from the Multidisciplinary Approach to the Study of Chronic Pelvic Pain (MAPP) Research Network (http://www.mappnetwork.org/).

Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis

Qi Long
University of Pennsylvania
E-mail: qlong@upenn.edu

Abstract: Multiple co-inertia analysis (mCIA) is a multivariate analysis method that can assess relationships and trends in multiple datasets. Recently it has been used for an integrative analysis of multiple highdimensional -omics datasets. However, the estimated loading vectors from the existing mCIA method are non-sparse, which presents challenges for interpreting analysis results. We propose two new mCIA methods: 1) a sparse mCIA (smCIA) method that produces sparse loading estimates and 2) a structured sparse mCIA (ssmCIA) method that further enables the incorporation of structural information among variables such as those from functional genomics. The two proposed methods achieve simultaneous model estimation and feature selection and yield analysis results that are more interpretable than the existing mCIA. Our extensive simulation studies demonstrate the superior performance of the smCIA and the ssmCIA methods compared to the existing mCIA. We also apply our methods to the integrative analysis of transcriptomics data and proteomics data from a cancer study.

An Unified Framework of Personalized Network Recovery and Detection

Ming Wang
Pennsylvania State University
E-mail: mwang@phs.psu.edu

Abstract: The Genotype-Tissue Expression Project (GTEx) aims to study how genes are differentially expressed across tissues to lead to human diseases. Whether and how human cells perform commoner unique functions across tissues are determined not only by genes individually but also through their co-expression network; however, inferring the tissue specificity of gene regulatory networks represents a substantial challenge. We address this challenge by developing a unified framework for assembling genomic data from multiple tissues into informative networks, importantly, adjusted for potential risk factors (e.g., gender, race). This framework utilize quasi-dynamic ordinary differential equations to recover tissue- and individual-specific gene networks with bi-directional, signed, and weighted interactions. This work provides a tool to compile, curate, and catalogue comprehensive encyclopedias for personalized gene regulatory networks, facilitating the translation of the GTEx data into clinical practice.

Estimation of personalized maximum tolerated dose (pMTD) by incorporation of patient's genomic profiles and all toxicity information in cancer Phase I clinical trial

Zhengjia Chen
Emory University
E-mail: zchen38@emory.edu

Abstract: Estimation of personalized maximum tolerated dose (pMTD) is the first critical step toward personalized medicine which can maximize the therapeutic effect of treatment for individual patient. To estimate pMTD, we propose to fully utilize the patient's biomarkers that can predict susceptibility to specific adverse events and response as covariates in a cutting-edge Bayesian adaptive and optimal cancer Phase I clinical trial design called EWOC-NETS. The methodology of incorporating patient's biomarker information in the estimation of pMTD for novel cancer therapeutic agent will be fully elaborated. Simulation studies demonstrate that utilization of biomarkers in EWOC-NETS can estimate pMTD while keeping its original merits: such as ethical constraint of overdose control and full utilization of all toxicity information to improve the accuracy and efficiency of pMTD estimation. A real cancer Phase I clinical trial will be presented to illustrate the utilization of genomics information for the estimation of pMTD.

S050: Novel Statistical Methods for Big Health Data

Clustering of Multivariate Data with Varying Dimensions

Bin Cheng
Columbia University
E-mail: bc2159@cumc.columbia.edu

Abstract: In many biomedical research, it is of interest to classify patients according to their medical information. Such information is not only of high but also varying dimensions. For example, the toxicity profile, microbial profile, or health record profile. We propose a method to cluster data of varying dimensions. Simulations demonstrate that the method has nice discriminating ability. A physical activity example is discussed as an illustration of the method.

An Adaptive Trial Design to Optimize Dose--Schedule Regimes with Delayed Outcomes

RUITAO LIN
The University of Texas MD Anderson Cancer Center

E-mail: ruitaolin@gmail.com

Abstract: We propose a two-stage phase I-II clinical trial design to optimize dose--schedule regimes of an experimental agent within ordered disease subgroups in terms of toxicity--efficacy tradeoff. The design is motivated by settings where prior biological information indicates it is certain that efficacy will improve with ordinal subgroup level. We formulate a flexible Bayesian hierarchical model to account for associations among subgroups and regimes, and to characterize ordered subgroup effects. Sequentially adaptive decision making is complicated by the problem, arising from the motivating application, that efficacy is scored on day 90 and toxicity is evaluated within 30 days from the start of therapy, while the patient accrual rate is fast relative to these outcome evaluation intervals. To deal with this in a practical way, we take a likelihood-based approach that treats unobserved toxicity and efficacy outcomes as missing values, and use elicited utilities that quantify the efficacy-toxicity trade-off as a decision criterion. Adaptive randomization is used to assign patients to regimes while accounting for subgroups, with randomization probabilities depending on the posterior predictive distributions of utilities. A simulation study is presented to evaluate the design's performance under a variety of scenarios, and to assess its sensitivity to the amount of missing data, the prior, and model misspecification.

Prediction of Alzheimer's disease by integrating local brain-network connectome

Yanming Li
University of Michigan
E-mail: liyanmin@umich.edu

Abstract: A novel approach for Alzheimer's disease (AD) prediction using brain-wide voxel-level imaging scans is presented. The proposed approach significantly improves the AD prediction accuracy by detecting and integrating the local predictive connectomic brain networks. A local predictive brain network contains not only marginally strong voxel signals, but also marginally weak signals in connection with the strong ones. Even though marginally weak signals by themselves exert no prediction effects, but they could exert strong prediction effect when in connection with the marginally strong signals. Marginally weak signals are usually ignored in the conventional brain-wide and voxel-level association or classification studies. The proposed approach detects both the marginally strong and weak signals and uncover their connected networks. The detected local brain network connectome provides biological insights on how the brain pathways attribute to AD development and prognosis.

We applied the approach to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The proposed approach achieves a prediction accuracy of 96.5%, much higher than that from other contemporary approaches without incorporating the marginally weak features. The proposed approach can be applied to prediction of other cognitive diseases or cancer subtypes using ultrahigh-dimensional imaging or genomic predictors.

S051: New Methodology for the Analysis of Neuroimaging Data

A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome] {A Bayesian Approach to Joint Modeling of Matrix-valued Imaging Data and Treatment Outcome with Applications to Depression Studies

Bei Jiang
University of Alberta
E-mail: bei1@ualberta.ca

Abstract: In this talk, we propose a unified Bayesian joint modeling framework for studying association between a binary treatment outcome and a baseline matrix-valued predictor. Specifically, a joint modeling approach relating an outcome to a matrix-valued predictor through a probabilistic formulation of multilinear principal component analysis (MPCA) is developed.

This framework establishes a theoretical relationship between the outcome and the matrix-valued predictor although the predictor is not explicitly expressed in the model.

Simulation studies are provided showing that the proposed method is superior or competitive to other methods, such as a two-stage approach and a classical principal component regression (PCR) in terms of both prediction accuracy and estimation of association; its advantage is most notable when the sample size is small and the dimensionality in the imaging covariate is large.

Finally, our proposed joint modeling approach is shown to be a very promising tool in an application exploring the association between baseline EEG data and a favorable response to treatment in a depression treatment study by achieving a substantial improvement in prediction accuracy in comparison to competing methods.

Improved prediction of brain age using multimodal neuroimaging data

Fengqing (Zoe) Zhang
Drexel University
E-mail: fz53@drexel.edu

Abstract: Prediction of brain age using neuroimaging data and machine learning has recently drawn increasing attention, as it has the potential to serve as a biomarker for characterizing the typical brain development and neuropsychiatric disorders. However, few studies examine multi-modal imaging features derived from MRI, DTI as well as rs-fMRI for brain age prediction. In addition, several studies report that the predicted brain age is often underestimated for older subjects and overestimated for younger subjects. We examine this systematic bias and propose different approaches to correct for the bias. We also compare different machine learning approaches to integrate different combinations of multi-modal imaging features. Furthermore, we apply our proposed approach to adolescents with anxiety disorders to test whether altered brain development is observed and how the brain development is related to cognitive deficits.

Multivariate Spline Estimation and Inference for Image-on-scalar Regression

Shan Yu
Iowa State University
E-mail: syu@iastate.edu

Abstract: Motivated by recent work of analyzing data in the biomedical imaging studies, we consider a class of image-on-scalar regression models for imaging responses and scalar predictors. We propose to use flexible multivariate splines over triangulations to handle the irregular domain of the objects of interest on the images and other characteristics of images. The proposed estimators of the coefficient functions are proved to be root-n consistent and asymptotically normal under some regularity conditions. We also provide a consistent and computationally efficient estimator of the covariance function. Asymptotic pointwise confidence intervals (PCIs) and

data-driven simultaneous confidence corridors (SCCs) for the coefficient functions are constructed. Our method can simultaneously estimate and make inferences of the coefficient functions while incorporating the spatial heterogeneity and spatial correlation. A highly efficient and scalable estimation algorithm is developed. Monte Carlo simulation studies are conducted to examine the finite-sample performance of the proposed method. The proposed method is applied to the spatially normalized Positron Emission Tomography (PET) data of Alzheimer's Disease Neuroimaging Initiative (ADNI).

Covariate Assisted Principal Regression for Covariance Matrix Outcomes with an Application to fMRI

Xi Luo
University of Texas Health Science Center at Houston
E-mail: xi.rossi.luo@gmail.com

Abstract: Modeling variances in data has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on vector covariates, collected from each observational unit. The main aim of this paper is to uncover the variation in the covariance matrices across units that are explained by the covariates. This paper introduces Covariate Assisted Principal (CAP) regression, an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the linear projections of the covariance matrices as well as the regression coefficients, and we establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic resonance imaging study, our approach identifies the human brain network changes associated with age and sex.

S052: Recent development of Gaussian approximation and its applications

Randomized incomplete U-statistics in high dimensions

Xiaohui Chen
University of Illinois at Urbana
E-mail: xiaohui@mit.edu

Abstract: This paper studies inference for the mean vector of a high-dimensional U-statistic. In the era of Big Data, the dimension d of the U-statistic and the sample size n of the observations tend to be both large, and the computation of the U-statistic is prohibitively demanding. Data-dependent inferential procedures such as the empirical bootstrap for U-statistics is even more computationally expensive. To overcome such computational bottleneck, incomplete U-statistics obtained by sampling fewer terms of the U-statistic are attractive alternatives. In this paper, we introduce randomized incomplete U-statistics with sparse weights whose computational cost can be made independent of the order of the U-statistic. We derive non-asymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions, namely in cases where the dimension d is possibly much larger than the sample size n, for both non-degenerate and degenerate kernels. In addition, we propose novel and generic bootstrap methods for the incomplete U-statistics that are computationally much less-demanding than existing bootstrap methods, and establish finite sample validity of the proposed bootstrap methods. The proposed bootstrap methods are illustrated on the application to nonparametric testing for the pairwise independence of a high-dimensional random vector under weaker assumptions than those appearing in the literature.

A Power One Test for Unit Roots Based on Sample Autocovariances

Guanghui Cheng
Guangzhou University
E-mail: chenggh845@nenu.edu.cn

Abstract: We propose a new unit-root test for a stationary null hypothesis H_0 against a unit-root alternative H_1. Our approach is nonparametric as the hull hypothesis only assumes that the process concerned is $I(0)$ without specifying any parametric forms. The new test is based on the fact that the sample autocovariance function (ACF) converges to the finite population ACF for an $I(0)$ process while it diverges to infinity with probability approaching one for a process with unit-roots. Therefore the new test rejects the null hypothesis for the large values of the sample ACF. To address the technical challenge `how large is large', we split the sample and establish an appropriate normal approximation for the null-distribution of the test statistic. The substantial discriminative power of the new test statistic is rooted from the fact that it takes finite value under H_0 and diverges to infinity with probability approaching one under H_1. This allows us to truncate the critical values of the test to make it with the asymptotic power one. It also alleviates the loss of power due to the sample-splitting. The finite sample properties of the test are illustrated by simulation which shows its stable and more powerful performance in comparison with the KPSS test \citep{kpss1992}.

Asymptotic mixed normality of realized covariance in high-dimensions

Yuta Koike
Graduate School of Mathematical Sciences, University of Tokyo
E-mail: kyuta@ms.u-tokyo.ac.jp

Abstract: The asymptotic mixed normality of the realized covariance matrix for a multi-dimensional continuous semimartingale observed at a high-frequency is established, where the dimension may be much larger than the sample size. More precisely, a mixed-normal approximation of the error distribution in terms of the Kolmogorov distance is shown in such a setting. The proof is based on a variant of the Chernozhukov-Chetverikov-Kato theory on high-dimensional central limit theorems for sums of independent random vectors, where the theory is accommodated to random asymptotic covariance matrices with the help of Malliavin calculus. Application to testing the residual sparsity of a continuous-time factor model is presented.

A Power One Test for Unit Roots Based on Sample Autocovariances

Jinyuan Chang
Southwestern University of Finance and Economics
E-mail: changjinyuan@swufe.edu.cn

Abstract: We propose a new unit-root test for a stationary null hypothesis H0against a unit-root alternative H1. Our approach is nonparametric as the hull hypothesis only assumes that the process concerned is I(0) without specifying any parametric forms. The new test is based on the fact that the sample autocovariance function (ACF) converges to the finite population ACF for an I(0) process while it diverges to infinity with probability approaching one for a process with unit-roots. Therefore the new test rejects the null hypothesis for the large values of the sample ACF. To address the

technical challenge 'how large is large', we split the sample and establish an appropriate normal approximation for the null-distribution of the test statistic. The substantial discriminative power of the new test statistic is rooted from the fact that it takes finite value under H0and diverges to infinity with probability approaching one under H1. This allows us to truncate the critical values of the test to make it with the asymptotic power one. It also alleviates the loss of power due to the sample-splitting. The finite sample properties of the test are illustrated by simulation which shows its stable and more powerful performance in comparison with the KPSS test (Kwiatkowski et al., 1992).

S053: New Advances in High-Dimensional Data Analysis

Extreme Quantile Estimation for Single Index Model

Deyuan Li
Fudan University
E-mail: deyuanli@fudan.edu.cn

Abstract: It is important to quantifying and predicting rare events which have huge effects. Existing work on analysing such effect mainly rely on either parametric model like linear quantile regression which lack of flexibility or non parametric model which subject to"the curse of dimensionality". We propose a new semi-parametric approach based on single index quantile regression. The proposed estimation are presented in three steps by first obtaining a root n-estimator of index parameter and then applying local polynomial regression to estimate the intermediate conditional quantiles which are then extrapolated to the tails. We establish asymptotic normality of the proposed estimator which balances better between model flexibility and parsimony. We also study its performance for finite sample by simulation and real data analysis to Los Angeles mortality rate, showing it is more accuate and stable than existing methods.

Quantiles, Expectiles and Jackknife Model Averaging in Ultra-High Dimensional Regressions

Yundong Tu
Peking University
E-mail: yundong.tu@gsm.pku.edu.cn

Abstract: Both quantile and expectile regressions are useful tools for modelling data with heterogeneous conditional distributions. This paper develops the Jackknife model averaging method for heteroskedastic quantile regressions and expectile regressions with ultra-high dimensional data. First, we propose an algorithm to screen all candidate variables and then select the relevant predictors for model averaging. In particular, we use the expectile partial correlation for screening in the expectile regression, in the spirit of the quantile partial correlation used for screening in the quantile regression (Ma, Li and Tsai, 2017). Theoretical results indicate that the screening procedure can achieve the sure screening set. Second, the model averaging expectile estimator using the leave-one-out cross-validated weight is shown to be asymptotically normal and asymptotically optimal in the sense of out-of-sample final prediction error. Numerical results demonstrate the nice performance of the screening procedure and the averaging estimators. Despite the fact that there exists a one-to-one mapping from expectiles to quantiles, it is found that the expectile-based model averaging estimator provides superior performance for estimating the conditional tail quantiles, as compared to the direct quantile-based approach (Lu and Su, 2015).

Testing Serial Correlation and ARCH Effect of High-Dimensional Time-Series Data

Yaxing Yang
Xiamen University
E-mail: yangyx@xmu.edu.cn

Abstract: This article proposes several tests for detecting serial correlation and ARCH effect in high-dimensional data. The dimension of data p = p(n) may go to infinity when the sample size n→∞. It is shown that the sample autocorrelations and the sample rank autocorrelations (Spearman's rank correlation) of the L1-norm of data are asymptotically normal. Two portmanteau tests based, respectively, on the norm and its rank are shown to be asymptotically $\chi 2$-distributed, and the corresponding weighted portmanteau tests are shown to be asymptotically distributed as a linear combination of independent $\chi 2$ random variables. These tests are dimension-free, that is, independent of p, and the norm rank-based portmanteau test and its weighted counterpart can be used for heavy-tailed time series. We further discuss two standardized norm-based tests. Simulation results show that the proposed test statistics have satisfactory sizes and are powerful even for the case of small n and large p. We apply the tests to two real datasets. Supplementary materials for this article are available online.

Targeted integrative learning with applications in suicide risk prediction

Kun Chen
University of Connecticut
E-mail: kun.chen@uconn.edu

Abstract: In many scientific problems, the goal is to make inference on a specified "target population" of interest. For example, in a single-arm clinical trial, the target population can be defined by the treated patients and the key is to find out what happens to them if they were not treated; in a suicide risk study, the target population may be consist of patients who received care from a specific healthcare provider. Yet, the data available may go way beyond the target population. As such, a crucial question is how to best integrate all the information to improve the inference for the target. In this talk, we consider two scenarios. For the scenario of "integrated data", we propose a distance-segmented regression (DSR) framework, in which a distance metric is used to measure how close each sample is to the target and is assumed to guide the conditional association between the outcome and predictors. For the scenario of "non-integratable" data, we propose a transfer learning model, in which the target population and the external database are linked through subject similarities. Applications in suicide risk prediction with medical claim data will be discussed.

S054: Statistical methodologies in clinical trials

Efficient Sample Size Adaptation Strategy With Adjustment Of Randomization Ratio

Yijie Zhou
Vertex Pharmaceuticals
E-mail: yijie_zhou@vrtx.com

Abstract: In clinical trials, sample size re-estimation is a useful strategy to mitigate the risk of uncertainty in design assumptions to ensure sufficient power for the final analysis. In current literature, sample size re-estimation and corresponding type I error control are discussed in the context of maintaining the original randomization ratio across treatment groups, which we refer to as "proportional increase." In practice, not all studies are designed based on an optimal randomization ratio due to practical reasons.

In such cases, when sample size is to be increased, it is more efficient to allocate the additional subjects such that the randomization ratio is brought closer to an optimal ratio. In this research, we propose an adaptive randomization ratio change when sample size increase is warranted. We refer to this strategy as "nonproportional increase," as the number of subjects increased in each treatment group is no longer proportional to the original randomization ratio. The proposed method boosts power not only through the increase of the sample size, but also via efficient allocation of the additional subjects. The control of type I error rate is shown analytically. Simulations are performed to illustrate the theoretical results.

Historical data borrowing from multiple historical trials

Bingzhi Zhang
Sanofi
E-mail: bingzhi.zhang@sanofi.com

Abstract: Innovative study design utilizing the information borrowed from external or historical data can greatly improve the efficiency of clinical trials by increasing power, reducing sample size and shortening study duration, leading to potentially a cost-effective trial design. In this presentation, several borrowing methods utilizing historical data from multiple sources will be demonstrated, and the frequentist operating characteristics of these methods will be shown.

Utilization of Robust Estimates of Treatment Effect via Semi-Parametric Models in MRCT

Ming Tan
Georgetown University
E-mail: mtt34@georgetown.edu

Abstract: Multi-Regional Clinical Trial (MRCT) plays an increasingly important role in global drug development. It serves as an efficient way to accelerate drug development, allowing global simultaneous development and earlier access to new drugs, benefiting patients worldwide. MRCTs provide an opportunity for health authorities to examine robustness and the applicability of a treatment across diverse populations while assessing country specific benefit risk profile. I will highlight some key statistical issues recognized in ICH E17, e.g., increased heterogeneity in trials involving different regions. In addition, despite randomization, there may be a differential treatment effect among different regions, potentially due to confounding region specific factors. Thus, accurate and robust estimates of variation would be especially important to MRCT. Most current methods for the assessment of the consistency or similarity of the treatment effect between different ethnic groups are based on some subjectively specified model. In this talk I will summarize recent advances on robust estimates of global and regional treatment effects in MRCT though semiparametric modeling, and show the asymptotic and finite sample properties of the estimate and how they are applied to real clinical trials. (This work is in collaboration with Ao Yuan, Chaojie Yang, Shuxin Wang, and Shuqi Wang.)

S055: Random Matrices Theory and Applications

Spectral graph matching and regularized quadratic relaxations

Zhou Fan
Yale University
E-mail: zhou.fan@yale.edu

Abstract: Given two unlabeled, edge-correlated graphs on the same set of vertices, we study the "graph matching" problem of matching the vertices of the first graph to those of the second. We propose a new spectral method for this problem, which first constructs a similarity matrix as a weighted sum of outer products between all pairs of eigenvectors of the two graphs, with weights given by a Cauchy kernel applied to the separation of the corresponding eigenvalues, then outputs a matching by a simple rounding procedure. The similarity matrix can also be interpreted as the solution to a regularized quadratic programming relaxation of the quadratic assignment problem. We show that for a correlated Erdos-Renyi model, this method returns the exact matching with high probability if the graphs differ by at most a 1/polylog(n) fraction of edges, both for dense graphs and for sparse graphs with at least polylog(n) average degree.

Conformal prediction with localization

Leying Guan
Yale University
E-mail: leying.guan@gmail.com

Abstract: In this paper, we propose the method of localized conformal prediction where we can perform the conformal inference using only a local region around a new test sample when constructing its confidence interval. The constructed confidence intervals have finite coverage guarantee for all underlying distributions. This is the first work that generalizes the method of conformal prediction to the case where we can break the data exchangeability and give the test sample a special role.

We provide theoretical results about its coverage guarantee and characterization of its over-overage, and we have applied it to different simulations and compared it with the method of conformal prediction.

Optimality of Spectral Clustering for Gaussian Mixture Model

Anderson Zhang
The Wharton School, University of Pennsylvania
E-mail: ayz@wharton.upenn.edu

Abstract: Spectral clustering is one of the most popular algorithms to group high dimensional data. It is easy to implement and computationally efficient. Despite its popularity and successful applications, its theoretical properties have not been fully understood. The spectral clustering algorithm is often used as a consistent initializer for more sophisticated clustering algorithms. However, in this paper, we show that spectral clustering is actually already optimal in the Gaussian Mixture Model, when the number of clusters of is fixed and consistent clustering is possible. Contrary to that spectral gap conditions are widely assumed in literature to analyze spectral clustering, these conditions are not needed in this paper to establish its optimality.

Adaptation in multivariate log-concave density estimation

Kyoung Hee Arlene Kim
Korea University
E-mail: arlenent@gmail.com

Abstract: We study the adaptation properties of the multivariate log-concave maximum likelihood estimator over two subclasses of log-concave densities. The first consists of densities with polyhedral support whose logarithms are piecewise affine. The complexity of such densities f can be measured in terms of the sum of the numbers of facets of the subdomains in the polyhedral subdivision of the support induced by f. Given n independent observations from a d-dimensional log-concave density with d=2 and d=3, we prove a sharp oracle inequality, which in particular implies that the Kullback-Leibler risk of the log-concave maximum likelihood estimator for such densities is bounded above by $\mathcal{O}(f)=n$, up to a polylogarithmic factor. Thus, the rate can be essentially parametric, even in this multivariate setting. The second type of subclass

consists of densities whose contours are well-separated; these new classes are constructed to be affine invariant and turn out to contain a wide variety of densities, including those that satisfy Holder regularity conditions. Here, we prove another sharp oracle inequality.

S056: Recent Advances on the Analysis of Failure Time Data

Penalized Generalized Empirical Likelihood with a Diverging Number of General Estimating Equations for Censored Data

Xingqiu Zhao
The Hong Kong Polytechnic University
E-mail: xingqiu.zhao@polyu.edu.hk

Abstract: This article considers simultaneous variable selection and parameter estimation as well as hypothesis testing in censored survival models without a parametric likelihood available. For the problem, we utilize certain growing dimensional general estimating equations and propose a penalized where the general estimating equations are constructed based on the semiparametric efficiency bound of estimation with given moment conditions. The proposed penalized generalized empirical likelihood estimators enjoy the oracle properties, and the estimator of any fixed dimensional vector of nonzero parameters achieves the semiparametric efficiency bound asymptotically. Furthermore, we show that the penalized generalized empirical likelihood ratio test statistic has an asymptotic central chi-square distribution. The conditions of local and restricted global optimality of weighted penalized generalized empirical likelihood estimators are also discussed. We present a two-layer iterative algorithm for efficient implementation, and investigate its convergence property. The performance of the proposed methods is demonstrated by extensive simulation studies, and a real data example is provided for illustration.

Empirical likelihood for additive hazards regression model with case II interval censored failure time data

Chunjie Wang
Changchun University of Technology
E-mail: cjwang2014@126.com

Abstract: Interval censored failure time data occur in many areas. Many approaches have been proposed under various hazards regression models based on the asymptotic normality in survival statistics studies. We proposed an empirical likelihood approach for an additive hazards model with case II interval censored failure time data. For a vector of regression parameters, an empirical log-likelihood ratio is defined and it is shown its limiting distribution is a standard chi-squared distribution. Finite sample performance of our proposed empirical likelihood approach are demonstrated by simulation studies, and it shows that the empirical likelihood method provides more accurate inference than the normal approximation method. Empirical likelihood appraoch is applied to analyzing a real study of the breast cancer data.

Semiparametric analysis of the additive hazards model with informatively interval-censored failure time data

Shuying Wang
Changchun University of Technology
E-mail: wangshuying0601@163.com

Abstract: Regression analysis of failure time data has been discussed by many authors and for this, one of the commonly used models is the additive hazards model, for which some inference procedures have been developed for various types of censored data. In this paper, a much general type of censored data, case K informatively interval-censored data, is considered for which there does not seem to exist an established inference procedure. For the problem, a joint modeling approach that involves a two-step estimation procedure and the sieve maximum likelihood estimation is presented. The proposed estimators of regression parameters are shown to be consistent and asymptotically normal, and a simulation study conducted suggests that the proposed procedure works well for practical situations. In addition, an application is provided.

Regression Analysis of Case-cohort Studies in the Presence of Dependent Interval Censoring

MINGYUE DU
Jilin University
E-mail: 1552091779@qq.com

Abstract: The case-cohort design is widely used as a means of reducing the cost in large cohort studies, especially when the disease rate is low and covariate measure ments may be expensive, and has been discussed by many authors. In this paper, we discuss regression analysis of case-cohort studies that produce interval-censored fail ure time with dependent censoring, a situation for which there does not seem to exist an established approach. For inference, a sieve inverse probability weighting estima tion procedure is developed with the use of Bernstein polynomials to approximate the unknown baseline cumulative hazard functions. The proposed estimators are shown to be consistent and the asymptotic normality of the resulting regression parameter estimators are established. A simulation study is conducted to assess the finite sam ple properties of the proposed approach and indicates that it works well in practical situations. The proposed method is applied to an HIV/AIDS case-cohort study that motivated this investigation.

S057: Statistical Analysis of Complex Data

Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression

Hui Zhao
Zhongnan University of Economics and Law
E-mail: hzhao@mail.ccnu.edu.cn

Abstract: The simultaneous estimation and variable selection for Cox model has been discussed by several authors (Fan and Li, 2002; Huang and Ma, 2010; Tibshirani, 1997) when one observes right-censored failure time data. However, there does not seem to exist an established procedure for interval-censored data, a more general and complex type of failure time data, except two parametric procedures in Scolas et al. (2016) and Wu and Cook (2015). To address this, we propose a broken adaptive ridge (BAR) regression procedure that combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage. In particular, the method allows for the number of covariates to be diverging with the sample size. Under some weak regularity conditions, unlike most of the existing variable selection methods, we establish both the oracle property and the grouping effect of the proposed BAR procedure. We conduct an extensive simulation study and show that the proposed approach works well in practical situations and deals with the collinearity problem better than the other oracle-like methods. An application is also provided.

A Vine Copula Approach for Regression Analysis of Bivariate Current Status Data with Informative Censoring

Huiqiong Li
Yunnan University
E-mail: lihuiqiong@ynu.edu.cn

Abstract: Bivariate current status data occur in many areas and many authors have discussed their analysis and proposed many inference procedures (Hu et al., 2017; Jewell et al., 2005; Wang et al., 2015). However, most of these methods are for the situation where the observation or censoring is non-informative and sometimes one may face informative censoring (Chen et al., 2012; Ma et al., 2015; Zhang et al., 2005), where one has to deal with three correlated random variables. In this paper, a vine copula approach is developed for regression analysis of bivariate current status data in the presence of informative censoring. The proposed estimators are shown to be strongly consistent and the asymptotic normality and efficiency of the estimated regression parameter are also established. Numerical results suggest that the proposed method works well in practice.

Personalized Glucose Prediction Using Attention-based RNN

Ran Duan
Eli Lilly and Company
E-mail: duan_ran@lilly.com

Abstract: Diabetes have been one of the leading cause of death in the US, which has taken a growing toll on people's health. Numerous treatments have been developed to battle diabetes, however, hypoglycemia, a condition of abnormally low level of blood glucose (<= 70mg/dL), is a common major adverse event for the diabetes management. Proactive prediction of patients glucose level, especially hypoglycemia event could greatly improve the adherence of insulin therapy and potentially promote the treatment effect. In this paper, we study the problem of blood glucose forecasting and provide a deep personalized solution. Our proposed method has several key advantages over existing methods: 1- it learns a personalized model for each patient as well as a global model; 2- it uses an attention mechanism and extracted time features to better learn long-term dependencies in the data; 3- it introduces a new, robust training procedure for time series data. We empirically show the efficacy of our model on a real dataset.

Regression analysis of informatively interval-censored failure time data with semiparametric linear transformation model

Da Xu
Center for Applied Statistical Research, School of Mathematics, Jilin University
E-mail: xuda302@126.com

Abstract: Regression analysis of interval-censored failure time data with noninformative censoring have been widely investigated and many methods have been proposed. Sometimes the mechanism behind the interval censoring may be informative and several approaches have also been developed for this latter situation. However, all of these existing methods are for single models and it is well-known that in many situations, one may prefer more flexible models. Corresponding to this, the linear transformation model is considered and a maximum likelihood estimation method is established.

In the proposed method, the association between the failure time of interest and the censoring time is modeled by the copula model, and the involved nonparametric functions are approximated by spline functions. The large sample properties of the proposed estimators are derived. Numerical results show that the proposed method performs well in practical application. Besides, a real data example is presented for the illustration.

S058: Limit Theorems of Random Fields and Related Topics

Some recent results on multivariate Gaussian random fields

Yimin Xiao
Michigan State Univeristy
E-mail: xiaoy@msu.edu

Abstract: In this talk, we present some recent results on probabilistic and statistical properties of a large class of multivariate Gaussian random fields with stationary increments including operator fractional Brownian motion and vector-valued operator-scaling random fields. The main results characterize properties of their local times, prediction errors, and coherence.

Derivatives of local times for some Gaussian fields

Fangju Xu
East China Normal University
E-mail: fjxu@finance.ecnu.edu.cn

Abstract: In this article, we consider derivatives of local time for a $(2,d)$-Gaussian field
[Z=big{ Z(t,s)= X^{H_1}_t -widetilde{X}^{H_2}_s, s,t ge 0big},] where X^{H_1} and $widetilde{X}^{H_2}$ are two independent processes from a class of d-dimensional centered Gaussian processes satisfying certain local nondeterminism property. We first give a condition for existence of derivatives of the local time. Then, under this condition, we show that derivatives of the local time are H"{o}lder continuous in both time and space variables. Moreover, under some additional assumptions, we show that this condition is also necessary for existence of derivatives of the local time at the origin.

Probabilities of deviations for record numbers in random walks

Yuqiang Li
School of statistics, East China Normal University
E-mail: yqli@stat.ecnu.edu.cn

Abstract: Record numbers, i.e. the numbers of ladder points, are basic statistics in random walks, whose deviation principles are unknown so far. In this talk, the asymptotic probabilities of moderate and small deviations for the record numbers in some random walks on line are introduced.

The moduli of non-differentiability for Gaussian random fields with stationary increments

Wensheng Wang
School of Economics, Hangzhou Dianzi University
E-mail: wswang@hdu.edu.cn

Abstract: We establish the exact moduli of non-differentiability of Gaussian random fields with stationary increments. As an application of the result, we prove that the uniform H¨older condition for the maximum local times of Gaussian random fields with stationary increments obtained in Xiao (1997) is optimal. These results are applicable to fractional Riesz-Bessel processes and stationary Gaussian random fields in the Mat´ern and Cauchy classes.

S059: Estimation from imperfectly observed data

Learning from EMR/EHR data to estimate treatment effects using high dimensional claims codes

Ronghui Xu
University of California
E-mail: rxu@ucsd.edu

Abstract: Our work was motivated by the analysis projects using the linked

US SEER-Medicare database to studying treatment effects in men of age 65 years or older who were diagnosed with prostate cancer. Such data sets contain up to 100,000 human subjects and over 20,000 claim codes. The data were obviously not randomized with regard to the treatment of interest, for example, radical prostatectomy versus conservative treatment. Informed by previous instrumental variable (IV) analysis, we know that confounding mostly likely exists beyond the commonly captured clinical variables in the database, and meanwhile the high dimensional claims codes have been shown to contain rich information about the patients' survival. Hence we aim to incorporate the high dimensional claims codes into the estimation of the treatment effect. The orthogonal score method is one that can be used for treatment effect estimation and inference assuming only consistency under the high dimensional hazards outcome model and the high dimensional conditional treatment model. In addition, we show that further refinement of the approach has doubly-robust properties in high dimensions: the resulting estimator is consistent when either of the hazards model or the treatment model is misspecified, as long as the other model is correct. We also develop a novel sparsity doubly robust result, where either the outcome or the treatment model can be a fully dense high-dimensional model.

Variable selection and estimation in generalized linear models with measurement error

Liqun Wang
University of Manitoba
E-mail: Liqun.Wang@umanitoba.ca

Abstract: We study the variable selection problem in linear and generalized linear models when some of the predictors are measured with error. We demonstrate how measurement error (ME) affects the selection results and propose regularized instrumental variable (RIV) methods to correct for the ME effects. We show that the proposed estimators have the oracle property in a linear model and we derive their asymptotic distribution under general conditions. We also investigate the performances of the estimators in generalized linear models. Our simulation studies show that the RIV estimators outperform the naive estimator in both linear and some generalized linear models. Finally, the proposed method is applied to a real dataset. This is a joint work with Lin Xue.

Estimating the covariance of fragmented functional data

Wei Huang
University of Melbourne
E-mail: wei.huang@unimelb.edu.au

Abstract: We consider the problem of estimating the covariance function of functional data which are only observed on a subset of their domain, such as fragments observed on small intervals or related types of functional data. We focus on situations where the data enable to compute the empirical covariance function or smooth versions of it only on a subset of its domain which contains a diagonal band. We show that estimating the covariance function consistently outside that subset is possible as long as the curves are suffciently smooth. We establish conditions under which the covariance function is identiable on its entire domain and propose a tensor product series approach for estimating it consistently. We derive asymptotic properties of our estimator and illustrate its nite sample properties on simulated and real data.

Density Estimation of Usual Intake for Food Consumption Data

Zhendong Huang
School of Mathematics and Statistics, The University of Melbourne
E-mail: huang.z@unimelb.edu.au

Abstract: In nutrition study, one of the research of interests is to estimate the distributions of individuals' usual intakes of episodically consumed foods, such as fruit, meat, alcohol, etc. However, the usual intakes can not be directly obtain, due to the heavy cost of nutritional surveys. Alternatively, contaminated version of the usual intakes (i.e. 24-hour dietary recalls) are observed, with significant measurement errors. Some foods, like alcohol, are never consumed by a proportion of people, making the usual intake a mixture of discrete and continuous distribution. This phenomenon makes existing non-parametric approaches break down, and new methods need to be developed for such data. We propose a new model regarding the food, which is consumed by only a proportion of population. A new estimation approach is developed and studied in terms of theoretical and numerical aspects.

S060: New approaches and modifications to modern computations

Ensemble Classification via Sufficient Dimension Reduction

Yingcun Xia
National University of Singapore
E-mail: staxyc@nus.edu.sg

Abstract: We propose an ensemble classification method for high-dimensional data by aggregating results of classifiers based on dimension reduction in randomly projected subspaces (DRIPS) of the features. We implement the method by the outer product gradients (OPG) method for dimension reduction and $k-$nearest neighbors (kNN) classifier. The method can select the tuning parameters, such as the efficient dimension and the number of neighbours for the classifier, efficiently. A voting method is proposed to aggregate the classification results for the final assignment of classification. The performance is compared with some of the popular classification methods using both simulated and many real data examples.

Zero-inflated negative-binomial NMF

Hiroyasu Abe
Kyoto University
E-mail: hiroabe@kuhp.kyoto-u.ac.jp

Abstract: co-authors are Hiroyasu Abe (Kyoto University, Kyoto, Japan) and Hiroshi Yadohisa (Doshisha, University, Kyoto, Japan) Nonnegative matrix factorization (NMF) is a matrix decomposition technique to capture hidden structures in a nonnegative data matrix, the entries of which are all nonnegative, for example, multivariate count data. The solution of NMF differs depending on what probability distribution is assumed. A Poisson distribution is the most commonly used probability distribution for modeling count data in NMF. Moreover, a new NMF method has been proposed recently based on the negative binomial distribution, which is compatible with overdispersed count data. However, these NMF methods do not account for the zero-inflated case, wherein the data matrix has many zeros. The zero-inflated Poisson distribution is a solution for the zero-inflated case, and NMF based on such a distribution has already been proposed. However, the NMF method to take into account zero-inflated case using the negative binomial distribution is not yet proposed. In this study, we propose the new NMF method using zero-inflated negative binomial distribution (ZINBNMF) to consider the overdispersion and zero-inflation of count data. We evaluate its performance by numerical

simulation.

Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage

Johan Lim
Seoul National University
E-mail: yohanlim@gmail.com

Abstract: We study the positive definiteness (PDness) problem in covariance matrix estimation. For high-dimensional data, many regularized estimators have been proposed under structural assumptions on the true covariance matrix, including sparsity. They were shown to be asymptotically consistent and rate-optimal in estimating the true covariance matrix and its structure. However, many of them do not take into account the PDness of the estimator and produce a non-PD estimate. To achieve PDness, researchers considered additional regularizations (or constraints) on eigenvalues, which make both the asymptotic analysis and computation much harder. In this paper, we propose a simple modification of the regularized covariance matrix estimator to make it PD while preserving the support. We revisit the idea of linear shrinkage and propose to take a convex combination between the first-stage estimator (the regularized covariance matrix without PDness) and a given form of diagonal matrix. The proposed modification, which we call the FSPD (Fixed Support and Positive Definiteness) estimator, is shown to preserve the asymptotic properties of the first-stage estimator if the shrinkage parameters are carefully selected. It has a closed form expression and its computation is optimization-free, unlike existing PD sparse estimators. In addition, the FSPD is generic in the sense that it can be applied to any non-PD matrix, including the precision matrix. The FSPD estimator is numerically compared with other sparse PD estimators to understand its finite-sample properties as well as its computational gain. It is also applied to two multivariate procedures relying on the covariance matrix estimator - the linear minimax classification problem and the Markowitz portfolio optimization problem - and is shown to improve substantially the performance of both procedures.

Generalized interventional approach for causal mediation analysis with causally ordered multiple mediators

Sheng-Hsuan Lin
National Chiao-Tung University (Taiwan), Institute of Statistics
E-mail: shenglin@stat.nctu.edu.tw

Abstract: Causal mediation analysis has demonstrated the advantage of mechanism investigation. In conditions with causally ordered mediators, path-specific effects (PSEs) are introduced for specifying the effect subject to a certain combination of mediators. However, most PSEs are unidentifiable. To address this, an alternative approach termed interventional analogue of PSE (iPSE), is widely applied to effect decomposition. Previous studies that have considered multiple mediators have mainly focused on two-mediator cases due to the complexity of the mediation formula. This study proposes a generalized interventional approach for the settings, with the arbitrary number of ordered multiple mediators to study the causal parameter identification as well as statistical estimation. It provides a general definition of iPSEs with a recursive formula, assumptions for nonparametric identification, a regression-based method, and a g-computation algorithm to estimate all iPSEs. We demonstrate that each iPSE reduces to the result of linear structural equation modeling subject to linear or log-linear models. This approach is applied to a Taiwanese cohort study for exploring the mechanism by which hepatitis C virus infection affects mortality through hepatitis B virus infection, liver function, and hepatocellular carcinoma. Software based on a g-computation algorithm allows users to easily apply this method for data analysis subject to various model choices according to the substantive knowledge for each variable. All methods and software proposed in this study contribute to comprehensively decompose a causal effect confirmed by data science and help disentangling causal mechanisms when the natural pathways are complicated.

S061: Advanced statistical modeling for complex data

New Tests for Equality of Several Covariance Functions for Functional Data

Jia Guo
Zhejiang University of Technology
E-mail: jia.guo@u.nus.edu

Abstract: In this article, we propose two new tests for the equality of the covariance functions of several functional populations, namely, a quasi-GPF test and a quasi-Fmax test whose test statistics are obtained via globalizing a pointwise quasi-F-test statistic with integration and taking its supremum over some time interval of interest, respectively. Unlike several existing tests, they are scale-invariant in the sense that their test statistics will not change if we multiply each of the observed functions by any nonzero function of time. We derive the asymptotic random expressions of the two tests under the null hypothesis and show that under some mild conditions, the asymptotic null distribution of the quasi-GPF test is a chi-squared-type mixture whose distribution can be well approximated by a simple-scaled chi-squared distribution. We also propose a random permutation method for approximating the null distributions of the quasi-GPF and Fmax tests. The asymptotic distributions of the two tests under a local alternative are also investigated and the two tests are shown to be root-n consistent. A theoretical power comparison between the quasi-GPF test and the L2-norm-based test proposed in the literature is also given. Simulation studies are presented to demonstrate the finite-sample performance of the new tests against five existing tests. An illustrative example is also presented.

Pairwise-rank-likelihood methods for the semiparametric transformation model

Tao Yu
National University of Singapore
E-mail: stayt@nus.edu.sg

Abstract: In this paper, we study the linear transformation model in the most general setup. This model includes many important and popular models in statistics and econometrics as special cases. Although it has been studied for many years, the methods in the literature either are based on kernel-smoothing techniques or make use of only the ranks of the responses in the estimation of the parametric components. The former approach needs a tuning parameter, which is not easily optimally specified in practice; and the latter approach may be {less accurate and computationally expensive}. In this paper, we propose a {pairwise rank likelihood} method {and extend it to a score-function-based method. Our methods estimate} all the unknown parameters in the linear transformation model, and we {explore the theoretical properties of} our proposed estimators. Via extensive numerical studies, we demonstrate that {our methods are} appealing in that the estimators are not only robust to the distribution of the random errors

but also {in many cases more accurate} than those of the existing methods.

Adaptive log-linear zero-inflated generalized Poisson autoregressive model with applications to crime counts

Xiaofei Xu
National University Of Singapore
E-mail: xu.xiaofei@u.nus.edu

Abstract: This research proposes a comprehensive ALG model (Adaptive Log-linear zero-inflated Generalized Poisson integer-valued GARCH) to describe the dynamics of integer-valued time series of criminal incidents with the features of autocorrelation, heteroscedasticity, over-dispersion, and excessive number of zero observations. The proposed ALG model captures time-varying nonlinear dependence and simultaneously incorporates the impact of exogenous variables in a unified modeling framework. We use an adaptive approach to automatically detect subsamples of local homogeneity, over which the time-dependent parameters are estimated through an adaptive Bayesian Markov chain Monte Carlo sampling scheme. A simulation study shows stable and accurate finite sample performances of the ALG model under various scenarios. When implemented with the crime incidents data in Byron, Australia, the ALG model delivers a persuasive estimation of the stochastic intensity of criminals and provides insightful interpretations on both the structural breaks of intensity and the temperature impacts on different criminal categories. The findings show that the temperature effect is insignificant to ``malicious damage to property'', ``breach bail conditions'', and ``arson'', yet is positively relevant for ``non-domestic violence related assault'', ``steal from person'', and ``liquor offenses''. This is a joint work with Cathy Chen, Ying Chen and Xiancheng Lin.

Community Detection on Social Network with Complex Attributes

Wanjie Wang
National University of Signapore
F-mail: staww@nus.edu.sg

Abstract: Observing a social network with the connection between nodes and the attributes of the nodes, how to identify the community label for each node? With the development of the theory in social network, this problem is of growing interest.

Currently, most methods assume the dimension of the attributes is not large. With the development of technology, the attributes associated with one node may be high dimensional, and only a few of them are related with the community label of interest. It will hurt the efficiency of the previous algorithms.

We proposed a new community detection method for the node with attributes that is high dimensional and sparse. With the product of the adjacency matrix and the attribute matrix, we apply the attributed-SCORE algorithm, to get reasonable community detection results. We apply it to the statistician citation network with the abstract of each paper as the attribute.

S062 Some developments on semiparametric regression models and panel data

Powerful Tests for Parent-of-Origin Effects at Quantitative Trait Loci on the X Chromosome

Wing Kam Fung
The University of Hong Kong
E-mail: wingfung@hku.hk

Abstract: Parent-of-origin effects, which describe an occurrence where the expression of a gene depends on its parental origin, are an important phenomenon in epigenetics. Statistical methods for detecting parent-of-origin effects on autosomes have been investigated for 20 years, but the development of statistical methods for detecting parent-of-origin effects on the X chromosome is relatively new. In the literature, a class of Q-XPAT-type tests are the only tests for the parent-of-origin effects for quantitative traits on the X chromosome. In this talk, we propose classes of tests to detect parent-of-origin effects for quantitative trait values on the X chromosome. The proposed tests can accommodate complete and incomplete nuclear families with any number of daughters. The simulation study shows that our proposed tests produce empirical type I error rates that are close to their respective nominal levels, as well as powers that are larger than those of the Q-XPAT-type tests. The proposed tests are applied to a real data set on Turner's syndrome,and the proposed tests give a more significant finding than the Q-C-XPAT test.

Subgroup Analysis of Linear Model With Measurement Error

Yang Bai
Shanghai University of Finance and Econimics
E-mail: statbyang@mail.shufe.edu.cn

Abstract: How to identify different subgroups in a heterogeneous population plays an important role in areas such as precision medicine, personalized goods and services. In real life, we usually can not obtain the exact values of the variables because of measurement error. How to estimate the model more accurately in the presence of measurement error is also a problem worth studying. Therefore, this paper simultaneously considers the subgroup analysis and measurement error. Under the framework of linear regression model, a new method is proposed to solve the subgroup analysis with measurement error. In this paper, the idea of constructing unbiased estimating equations with two replicate measurements is transformed into minimizing an objective function and then concave penalty is applied to pairwise differences of the coefficients in order to estimate the coefficients and identify the subgroups simultaneously. This paper develops an alternating direction method of multipliers algorithm with concave penalties and demonstrate its convergence. The proposed estimators are proved to be of consistency and asymptotic normality, which are also supported by simulation. Finally, we apply our method to the data from the Lifestyle Education for Activity and Nutrition study.

High-dimensional expectile regression with a possible change point

YONG HE
Shandong University of Finance and Economics
E-mail: heyong@sdufe.edu.cn

Abstract: Large-dimensional factor model has drawn much attention in the big-data era, which characterizes the dependency structure of big-data set by a few latent factors and thus achieves great dimension reduction. Conventional methods for estimating factor model often ignore the effect of heavy-tailedness of data and thus may result in inefficient or even inconsistent estimation. In this paper, we propose robust estimators for both the factor loadings and factor scores by adopting the Elliptical Factor Model (EFM) framework. The robustness is achieved by a two-step estimation procedure. In the first step, Multivariate Kendall's tau matrix is employed to estimate the space spanned by the columns of the factor loading matrix. In the second step, we propose to estimate the factor scores by Ordinary Least

Square (OLS) regression. Theoretically, we show that the factor loadings and scores as well as the common components can be estimated consistently without exerting any moment condition. The finite sample performance of the proposed method is assessed through simulation and the analysis of a macroeconomic dataset.

S063: Recent advances in Bayesian analysis of complex data

Distributed Bayesian Inference for Varying Coefficient Spatiotemporal Models

Cheng Li
National University of Singapore
E-mail: stalic@nus.edu.sg

Abstract: Bayesian varying coefficient models based on Gaussian processes are popular in many disciplines because they balance flexibility and interpretability. Markov chain Monte Carlo (MCMC) methods are available to fit these models, but they are inefficient even for moderately large data. Motivated by the task of modeling massive spatiotemporal data, we develop a divide-and-conquer Bayesian method for fitting spatiotemporal varying coefficient models based on multiple output Gaussian processes. Our method partitions the space-time tuples into a large number of overlapping subsets, obtains MCMC samples of parameters and predictions in parallel across the subsets, and combines the subset MCMC samples into an approximate full data posterior. By tuning the stochastic approximation in subset posteriors, we show theoretically that the combined posterior distribution can converge at an optimal rate towards the true underlying surface, and we provide guidance for choosing the number of subsets depending on the analytic properties of Gaussian processes. To improve the efficiency of MCMC sampling, we further develop a new data augmentation scheme based on parameter expansion. We demonstrate the excellent empirical performance of our method across diverse simulations and a real data application to the temperature and precipitation data in the U.S.A.

Data assimilation from a viewpoint of regularization theory

Shuai Lu
Fudan University
E-mail: slu@fudan.edu.cn

Abstract: Inverse problems are ubiquitous in real applications. Understanding of algorithms for their solution has been greatly enhanced by a deep understanding of the linear inverse problem. In the applied communities ensemble-based filtering methods have recently been used to solve inverse problems by introducing an artificial dynamical system. This opens up the possibility of using a range of other filtering methods, such as 3DVAR, Kalman filter (online) and 4DVAR (offline), to solve inverse problems, again by introducing an artificial dynamical system. The aim of this talk is to undertand these methods in the context of the regularization theory under the framework of linear inverse problems.

Exploiting sparse conditional structure in MALA-within-Gibbs

Xin Tong
National University of Singapore
E-mail: mattxin@nus.edu.sg

Abstract: Markov chain Monte Carlo (MCMC) samplers are numerical methods for drawing samples from a given, targeted probability distribution. We discuss one particular MCMC sampler, the MALA-within-Gibbs sampler, from theoretical and practical perspectives. We first show that the acceptance rate and step size of this sampler are independent of the overall problem dimension when (1) the target distribution has sparse conditional structure; and (2) if this structure is reflected in the partially updating strategy of MALA-within-Gibbs. If, in addition, the target distribution is also block-wise log-concave, then the sampler's convergence rate is dimension independent. From a practical perspective, we expect that MALA-within-Gibbs is useful for solving high-dimensional Bayesian estimation where we expect sparse conditional structure to occur in the posterior distributions of many practically relevant problems. In this context, a partitioning of the state that correctly reflects the sparse conditional structure must be found and we illustrate this process in two numerical examples.

Accelerating Metropolis-within-Gibbs sampler with localized computations of differential equations

Qiang Liu
National University of Singapore
E-mail: matliuq@nus.edu.sg

Abstract: Inverse problem is ubiquitous in science and engineering, and Bayesian methodologies are often used to infer the underlying parameters. For high dimensional temporal-spatial models, classical Markov chain Monte Carlo (MCMC) methods are often slow to converge, and it is necessary to apply Metropolis-within-Gibbs (MwG) sampling on parameter blocks. However, the computation cost of each MwG iteration is typically O(n2), where n is the model dimension. This can be too expensive in practice. This paper introduces a new reduced computation methods to bring down the computation cost to O(n), for the inverse initial value problem of a stochastic differential equation (SDE) with local interactions. The key observation is that each MwG proposal is only different from the original iterate at one parameter block, and this difference will only propagate within a local domain in the SDE computations. Therefore we can approximate the global SDE computation with a surrogate updated only within the local domain for reduced computation cost. Both theoretically and numerically, we show that the approximation errors can be controlled by the local domain size. We discuss how to implement the local computation scheme using Euler-Maruyama and 4th order Runge-Kutta methods. We numerically demonstrate the performance of the proposed method with the Lorenz 96 model and a linear stochastic flow model.

S064: New methods of testing and classification in complex data

Innovated power enhancement for testing multi-factor pricing models with a large number of assets

Xiufan Yu
Pennsylvania State University
E-mail: xzy22@psu.edu

Abstract: Testing multi-factor pricing models with a large number of assets is instrumental for asset pricing theory and practice. Due to the accumulation of errors in estimating high-dimensional parameters, traditional quadratic-form tests such as the Wald test perform poorly against the sparse alternative hypothesis in the presence of a few mis-priced assets. Fan et al. (2015) introduced a powerful testing procedure by adding a power enhancement component to the Wald test statistic and proved the power enhancement properties. To provide a promising alternative to Fan et al. (2015), we first introduce a new maximum-form test statistic and then study the asymptotic joint distribution of the Wald test statistic and the maximum

test statistic. Surprisingly, we prove that two test statistics are asymptotically independent. Given their asymptotic independence, we propose an innovated power enhancement testing procedure to combine their respective power based on the celebrated Fisher's method (Fisher, 1925). Theoretically, we prove that the innovated power enhancement test retains the desired nominal significance level and achieves the asymptotically consistent power against the more general alternative. Furthermore, we demonstrate the nite-sample performance of our proposed innovated power enhancement test in simulations and a real application to testing market efficiency using asset returns of the Russel-2000 portfolio.

Topics on multiple testing

Hongyuan Cao
Florida State University
E-mail: hcao@fsu.edu

Abstract: High throughput technologies enable simultaneous inference of complex high dimensional data. An acute problem is the multiple testing adjustment. Most existing literature examine the problem under independence and sparsity assumptions. We propose a multiple testing procedure to incorporate dependence and non-sparsity features inherent in many high dimensional data, such as microRNA in genomics and quantitative high throughput screening (qHTS) assays in toxicology.

Hierarchical Community Detection with Fiedler Vectors

Xiaodong Li
UC Davis
E-mail: xdgli@ucdavis.edu

Abstract: Hierarchical clustering of entities based on observations of their connections has already been widely studied and implemented in the practice of network analysis. However, the statistical properties of diverse hierarchical community detection are still majorly unclear. We here study the binary tree stochastic block model in the literature to accommodate general compositions of edge probabilities. It can be shown that the eigen-structrues of the graph Laplacian of the population binary tree stochastic block model reveals the latent structure of the network at all levels. This fact inspires us to retrieve the hidden hierarchical structure of communities by using a recursive bi-partitioning algorithm with Fiedler vector, dividing a network into two communities repeatedly until a stopping rule indicates that there are no further communities. The method is further theoretically justified in sparse networks with the help of the newly developed theory about entrywise bound for eigenvector perturbations. The is based on an ongoing project with my student Xingmei Lou.

Principal Boundary on Riemannian Manifolds and Classification Problem

Zhigang Yao
National University of Singapore
E-mail: zhigang.yao@nus.edu.sg

Abstract: We will discuss the problem of finding principal components to the multivariate datasets, that lie on an embedded nonlinear Riemannian manifold within the higher-dimensional space. Our aim is to extend the geometric interpretation of PCA, while being able to capture the non-geodesic form of variation in the data. We introduce the concept of a principal sub-manifold, a manifold passing through the center of the data, and at any point of the manifold, it moves in the direction of the highest curvature in the space spanned by the eigenvectors of the local tangent space PCA. We show the principal sub-manifold yields the usual principal components in Euclidean space. We illustrate how to find, use and interpret the principal sub-manifold, with which a classification boundary can be defined for data sets on manifolds.

S065: Recent Advances in Statistical Theories and Applications

Estimation for Double-Nonlinear Cointegration

Qiwei Yao
London School of Economics
E-mail: q.yao@lse.ac.uk

Abstract: In recent years statistical inference for nonlinear cointegration has attracted attention from both academics and practitioners. This paper proposes a new type of cointegration in the sense that two univariate time series y(t) and x(t) are cointegrated via two (unknown) smooth nonlinear transformations. More precisely, it holds that G(y(t), b) =g(x(t))+u(t), where $G(., beta)$ is strictly increasing and known upto an unknown parameter b, $g(.)$ is unknown and smooth, x(t) is I(1), and $u(t)$ is the stationary disturbance. This setting nests the nonlinear cointegration model of Wang and Phillips (2009) as a special case with G(y, b)=y. It extends the model of Lindon et al (2008) to the cases with a unit-root nonstationary regressor. Sieve approximations to the smooth nonparametric function g are applied, leading to an extremum estimator for b and a plugging-in estimator for g(.). Asymptotic properties of the estimators are established, revealing that both the convergence rates and the limiting distributions depend intimately on the properties of the two nonlinear transformation functions. Simulation studies demonstrate that the estimators perform well even with small samples. A real data example on the environmental Kuznets curve portraying the nonlinear impact of per-capita GDP on air-pollution illustrates the practical relevance of the proposed double-nonlinear cointegration
(Joint work with Yingqian Lin and Yundong Tu.)

High-dimensional log-concave density estimation

Richard Samworth
University of Cambridge
E-mail: r.samworth@statslab.cam.ac.uk

Abstract: We tackle the problem of high-dimensional nonparametric density estimation by taking the class of log-concave densities on $\mathbb{R}^p$ and incorporating within it symmetry assumptions, which facilitate scalable estimation algorithms and can mitigate the curse of dimensionality. Our main symmetry assumption is that the super-level sets of the density are K-homothetic (i.e. scalar multiples of a convex body $K \subseteq \mathbb{R}^p$). When K is known, we prove that the K-homothetic log-concave maximum likelihood estimator based on n independent observations from such a density has a worst-case risk bound with respect to, e.g., squared Hellinger loss, of $O(n^{-4/5})$, independent of p. Moreover, we show that the estimator is adaptive in the sense that if the data generating density admits a special form, then a nearly parametric rate may be attained. We also provide worst-case and adaptive risk bounds in cases where K is only known up to a positive definite transformation, and where it is completely unknown and must be estimated nonparametrically.

Generative Link Prediction for Incomplete Networks with Node Features

Ji Zhu
University of Michigan

E-mail: jizhu@umich.edu

Abstract: Link prediction is one of the fundamental problems in network analysis. Most existing methods require at least partial observation of connections for every node. In real-world networks, however, there often exist nodes that do not have any link information, and it is imperative to make link predictions for such nodes based on their node features. In this talk, we consider a general framework in which a network consists of three types of nodes: nodes having features only, nodes having link information only, and nodes having both. Our goal is to predict links between nodes having features only and all other nodes. Under this setting, we have proposed a family of generative models for incomplete networks and node features, and we have developed a variational auto-encoder algorithm for model estimation and link prediction and investigated different encoder structures. We have also designed a cross-validation scheme under the problem setting. The proposed method has been evaluated on an online social network and two citation networks and achieved superior performance comparing with existing methods. This talk is based on joint work with Boang Liu, Binghui Liu and Elizaveta Levina.

Fourier Transform Approach for Inverse Dimension Reduction Method

Xiangrong Yin
University of Kentucky
E-mail: yinxiangrong@uky.edu

Abstract: Estimating an inverse regression space is especially important in sufficient dimension reduction. However, it typically requires a tuning parameter, such as the number of slices in a slicing method or bandwidth selection in a kernel estimation approach. Such a requirement not only affects the accuracy of estimates in a finite sample, but also increases difficulties for multivariate models. In this paper, we use a Fourier transform approach to avoid such difficulties and incorporate multivariate models. We further develop a Fourier transform approach to deal with variable selection, categorical predictor variables, and large p, small n data. To test the dimension, asymptotic results are obtained. Simulation studies and data analysis show the efficacy of our proposed methods.

S066: Recent Advances in Statistical Learning

Adaptive Design of Network A/B tests

Feifang Hu
George Washington University
E-mail: feifang@email.gwu.edu

Abstract: Controlled experiments (A/B tests) are currently very popular in many industries. Many controlled experiments have three common features: (i) try to estimate ATE (average treatment effect); (ii) usually sequential; and (iii) depending on important covariates. In many applications (both online experiment and clinical trials), the responses of subjects could be a mixture of treatment effect, network effect, spill-over effect and their covariates. We should consider both covariates and the network connection in both the design and analyze of these experiments. In this talk, we propose new adaptive designs for network A/B tests. We aim to minimize the mean squared error of the estimated difference of treatment effects, which is equivalent to improve network connection balance across treatment groups. Under mild assumption, we prove that the new procedure has smaller mean square error than complete randomization. The advantages of the proposed designs are also demonstrated by numerical studies.

Support vector machine in construction of personal treatment rules

Peter Song
University of Michigan
E-mail: pxsong@umich.edu

Abstract: One central task of personalized medicine is to establish individualized treatment rules (ITRs) for patients with heterogeneous responses to different treatments. Motivated from a diabetes clinical trial, we consider a problem of great relevance to translational medicine, where many biomarkers are potentially useful to improve an existing ITR. This calls for a screening procedure to assess added values of new biomarkers to derive an improved ITR. We propose net benefit index (NBI) that quantifies gain or loss of treatment benefit due to reclassification in which the optimal labels are obtained by support vector machine (SVM) in the context of outcome weighted learning (OWL). We calculate p-value of the proposed NBI-based test using the bootstrap null distribution generated by stratified permutations on individual treatment arm. The performance of the proposed method is evaluated by simulations and the motivating clinical trial. Our results show that the NBI-based test controls false discovery rate well and achieves high sensitivity. In addition, this screening method demonstrates an improved correct classification rate when ITR is expanded by including selected biomarkers. This is a joint work with Yiwang Zhou and Haodo Fu.

Random projection pursuit regression for high-dimensional complex data

Sijian Wang
Rutgers University
E-mail: sijian.wang@stat.rutgers.edu

Abstract: Projection pursuit regression (PPR) adapts the additive models in that it first projects the data matrix of explanatoryvariables in the optimal direction before applying smoothing functions to these explanatory variables. As a consequenceof this, PPR can be quite flexible to approximate a complicated regression function. It also has connections to neuronnetwork (deep learning) and boosting. One possible limitation of PPR is the complicated and heavy computation,especially when the number of variables is large. In this talk, borrowing the spirit of random forest, we introduce a random projection pursuit regression, which is not only as flexible as PPR, but also has an advantage in computingcomplexity. The method is demonstrated with both simulation studies and real data analysis. The connections to neuronnetwork and boosting will be also discussed.

Ultrahigh Dimensional Precision Matrix Estimation via Refitted Cross Validation

Zhao Chen
Fudan University
E-mail: zchen_fdu@fudan.edu.cn

Abstract: This paper develops a new estimation procedure for ultrahigh dimensional sparse precision matrix, the inverse of covariance matrix. Regularization methods have been proposed for sparse precision matrix estimation, but they may not perform well with ultrahigh dimensional data due to the spurious correlation. We propose a refitted cross validation (RCV) method for sparse precision matrix estimation based on its Cholesky decomposition, which does not require the Gaussian assumption. The proposed RCV procedure can be easily implemented with existing software for ultrahigh dimensional linear regression. We establish the consistency of the proposed RCV estimation and show that the rate of convergence of the RCV estimation without assuming banded structure is the same as that of

those assuming the banded structure in Bickel and Levina (2008b). Monte Carlo studies were conducted to access the finite sample performance of the RCV estimation. Our numerical comparison shows that the RCV estimation outperforms the existing ones in various scenarios. We further apply the RCV estimation for an empirical analysis of asset allocation.

S067: High dimensional statistical inference

Inter-subject correlation analysis with fMRI data

Hongnan Wang
University of Illinois at Chicago
E-mail: hwang313@uic.edu

Abstract: A focus of inter-subject correlation (ISC) analysis is to identify brain regions that respond similarly or synchronize to the same stimuli among a group of individuals by quantifying the inter-subject correlations. Functional MRI data are ideal for evaluating inter-subject correlation with continuous stimuli. In this talk, I will introduce a nonparametric test procedure that is valid under individual and temporal heterogeneity. We study the asymptotic distributions of the proposed test statistics for both random and fixed designs. Our empirical studies demonstrate that the proposed test procedure performs better than the commonly used methods in ISC studies, the adjusted Lagrange multiplier test, Pesaran's cross-sectional dependence (CD) test, and the adjusted Pesaran's CD test.

Individual Data Protected Integrative Regression Analysis of High-dimensional Heterogeneous Data

Yin Xia
Fudan University
E-mail: xiayin@fudan.edu.cn

Abstract: Evidence based decision making often relies on meta-analyzing multiple studies, which enables more precise estimation and investigation of generalizability. Integrative analysis of multiple heterogeneous studies is, however, highly challenging in the high dimensional setting. The challenge is even more pronounced when the individual level data cannot be shared across studies due to privacy concerns. Under ultra high dimensional sparse regression models and the constraint of not sharing individual data across studies, we propose in this paper a novel integrative estimation procedure by Aggregating and Debiasing Local Estimators (ADeLE). The ADeLE procedure protects individual data through summary-statistics-based integrating procedure, accommodates between study heterogeneity in both the covariate distribution and model parameters, and attains consistent variable selection. Furthermore, the prediction and estimation errors incurred by aggregating derived data is negligible compared to the statistical minimax rate. In addition, the ADeLE estimator is shown to be asymptotically equivalent in prediction and estimation to the ideal estimator obtained by sharing all data. The finite-sample performance of the ADeLE procedure is studied via extensive simulations. We further illustrate the utility of the ADeLE procedure to derive phenotyping algorithms for coronary artery disease using electronic health records data from multiple disease cohorts.

Focused Generalized Method of Moments for High-Dimensional Causal Structural Learning

Changcheng Li
Penn State University
E-mail: cxl508@psu.edu

Abstract: We propose a new constraint-based causal structural learning algorithm for high-dimensional Gaussian linear causal graphical models. Existing constraint-based approaches like the PC algorithm remove edges between vertices by carrying conditional independence tests on all possible candidates of d-separation sets. This can be computationally expensive and have exponential worst-case complexity. To tackle these issues, we propose a regularized approach called Focused Generalized Method of Moments (FGMM) to identify d-separation sets between vertices in this paper. Regularized approaches have been used to identify Markov blankets in causal graphical models. However, Markov blankets contain spouses besides true neighbors, which also need to be removed by searching d-separation sets. Distinguished from existing regularized approaches, the FGMM approach utilizes the moment conditions to identify d-separation sets directly. We further propose skeleton and structural learning algorithms based on the FGMM method, and establish the consistency of the FGMM algorithm in high-dimensional settings. We further conduct Monte Carlo simulations on various benchmark networks and show advantages of the proposed FGMM algorithm both in accuracy and speed.

S068: Large dimensional random matrix theory and its applications

Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression

Cheng Wang
Shanghai Jiao Tong University
E-mail: chengwang@sjtu.edu.cn

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

Beta matrix and testing the equality of two high dimensional covariance matrices

Jiang Hu
Northeast Normal University
E-mail: huj156@nenu.edu.cn

Abstract: In this talk, we will report some results about testing the equality of two high dimensional covariance matrices which are based on the Beta matrix, including the likelihood ratio tests, Pillai's trace tests and their asymptotic power functions.

The limits of the distant sample spikes for a high-dimensional generalized Fisher matrix and its applications

Dandan Jiang

Xi'an Jiaotong University
E-mail: jddpjy@163.com

Abstract: A generalized spiked Fisher matrix is considered in this paper, which is assumed to have a general form with the spiked eigenvalues scattered into a few bulks. By removing the diagonal or diagonal block-wise assumption in the previous works, we provide the limits of the distant sample spikes for a generalized Fisher matrix. To make it more applicable, we also give the estimates of the limits of the distant sample spikes, as well as the ones of the population spikes. As shown in the simulations, the proposed estimators are suitable for a wider range and more feasible in practice.

Community Detection Based on the L_infty convergence of eigenvectors in DCBM

Yan Liu
School of Mathematics and Statistics, Northeast Normal University
E-mail: liuy035@nenu.edu.cn

Abstract: Spectral clustering is one of the most popular algorithms for community detection in network analysis. Based on this rationale, in this paper we give the convergence rate of eigenvectors for the adjacency matrix in the l_infty norm, under the stochastic block model (BM) and degree corrected stochastic block model (DCBM), adding some mild and rational conditions. We also extend this result to a more general model, presented based on the DCBM such that the value of random variables in the adjacency matrix is not 0 or 1, but an arbitrary real number. During the process of proving the above conclusion, we obtain the relationship of the eigenvalues in the adjacency matrix and the corresponding `population' matrix, which vary in dimension from the community-wise edge probability matrix. Using that result, we can give an estimate of the number of the communities in a known set of network data. %to solve the problem that how to determine.

Meanwhile we proved the consistency of the estimator. Furthermore, according to the derivation of proof for the convergence of eigenvectors, we propose a new approach to community detection -- Spectral Clustering based on Difference of Ratios of Eigenvectors (SCDRE). Our simulation experiments demonstrate the superiority of our method in community detection.

S070: Recent advances on precision medicine and biomarker research

Constructing personalized decision algorithm for mHealth applications

Min Qian
Columbia University
E-mail: mq2158@cumc.columbia.edu

Abstract: Mental illnesses affect tens of millions of people each year. However, only half of those in need actually receive treatment. This is partly due to the substantial barriers associated with accessing office-based mental health care. As such, there are great needs for providing those who are in need of help with access to efficacious therapies. The use of mobile applications can fill the gap by delivering personalized treatments to patients who will otherwise not have access to the traditional treatments. In this work, we proposed a new analytical framework to develop personalized mobile decision algorithms to optimize immediate goals. The method is evaluated using simulation studies and illustrated using data from a recent mobile health study.

Improved doubly robust estimation in learning optimal individualized treatment rules

Yingqi Zhao
Fred Hutchinson Cancer Research Center
E-mail: yqzhao@fredhutch.org

Abstract: Individualized treatment rules (ITRs) recommend treatment according to patient characteristics. There is a growing interest in developing novel and ecient statistical methods in constructing ITRs. We propose an improved doubly robust estimator of the optimal ITRs. The proposed estimator is based on a direct optimization of an augmented inverse-probability weighted estimator (AIPWE) of the expected clinical outcome over a class of ITRs. The method enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent when either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among the class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. Individualized treatment rules (ITRs) recommend treatment according to patient characteristics. There is a growing interest in developing novel and ecient statistical methods in constructing ITRs. We propose an improved doubly robust estimator of the optimal ITRs. The proposed estimator is based on a direct optimization of an augmented inverse-probability weighted estimator (AIPWE) of the expected clinical outcome over a class of ITRs. The method enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent when either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among the class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model.

Optimizing personalized intervention from the aspect of health economics

Shuai Chen
University of California
E-mail: shschen@ucdavis.edu

Abstract: It is widely recognized that treatments often have substantially different effects across a population. Many statistical methods have recently been developed for identifying subgroups of patients who may benefit from different available treatments. Cost-effectiveness analysis (CEA) is an important component of the economic evaluation of new treatment options. In many clinical studies of costs, censored data pose challenges to the CEA. Due to the induced dependent censoring problem, standard survival analysis techniques are often invalid for censored costs. We propose a method for estimating individualized treatment benefits and costs with censored data, which would provide a tool for physicians and patients to make decisions based on personal characteristics and preference on benefit-cost tradeoff. Our method bypasses the modelling of main effect, and hence involves minimum modeling for the relationship between the outcome and covariates pertinent to measuring individual treatment benefit-cost tradeoff. We then conducted numerical studies to evaluate the performance of proposed method.

Learning Individualized Treatment Rules from Electronic Health Records

Yuanjia Wang
Columbia University
E-mail: yw2016@cumc.columbia.edu

Abstract: To address substantial heterogeneity in patient re- sponse to treatment of chronic disorders and achieve the promise of precision medicine, individualized treatment rules (ITRs) are estimated to tailor treatments according to patient-specific characteristics. Randomized controlled trials (RCTs) provide gold standard data for learning ITRs not subject to confounding bias. However, RCTs are often conducted under stringent inclu- sion/exclusion criteria, and participants in RCTs may not reflect the general patient population. Thus, ITRs learned from RCTs lack generalizability to the broader real world patient population. Real world databases such as electronic health records (EHRs) provide new resources as complements to RCTs to facilitate evidence-based research for personalized medicine. However, to ensure the validity of ITRs learned from EHRs, a number of challenges including confounding bias and selection bias must be addressed. In this work, we propose a matching-based machine learning method to estimate optimal individualized treatment rules from EHRs using interpretable features extracted from EHR documentation of medications and ICD diagnoses codes. We use a latent Dirichlet allocation (LDA) model to extract latent topics and weights as features for learning ITRs. Our method achieves confounding reduction in observational stud- ies through matching treated and untreated individuals and improves treatment optimization by augmenting feature space with clinically meaningful LDA-based features. We apply the method to EHR data collected at New York Presbyterian Hospital clinical data warehouse in studying optimal second-line treatment for type 2 diabetes (T2D) patients. We use cross validation to show that ITRs outperforms uniform treatment strategies (i.e., assigning same treatment to all individuals), and including topic modeling features leads to more reduction of post-treatment complications.

S071: New Advances of Adaptive Data Collection

Statistical Inference for Covariate-Adaptive RandomizationProcedures

Wei Ma
Renmin University of China
E-mail: mawei@ruc.edu.cn

Abstract: Covariate-adaptive randomization (CAR) procedures are frequently used in comparative studies to increase the covariate balance across treatment groups. However, because randomization inevitably uses the covariate information when forming balanced treatment
groups, the validity of classical statistical methods after such randomization is often unclear. In this article, we derive the theoretical properties of statistical methods based on general CAR under the linear model framework. More importantly, we explicitly unveil the relationship between covariate-adaptive and inference properties by deriving the asymptotic representations of the corresponding estimators. We apply the proposed general theory to various randomization procedures such as complete randomization, rerandomization, pairwise sequential randomization, and Atkinson's DA-biased coin design and compare their performance analytically. Based on the theoretical results, we then propose a new approach to obtain valid and more powerful tests. These results open a door to understand and analyze experiments based on CAR. Simulation studies provide further evidence of the advantages of the proposed framework and the theoretical results.

Randomization-Based Inference Following Randomized Clinical Trials

William Rosenberger
George Mason University
E-mail: wrosenbe@gmu.edu

Abstract: Experiments rely on replication rather than sampling from a population for their scientific validity. It was recognized by the pioneers of statistics that incorporating randomization into an experiment allows a basis for inference that cannot be obtained otherwise. Nonetheless, the advent of Neyman-Pearson inference led to inference based on random sampling becoming the standard method for analyzing randomized clinical trials. The principle reason for this anomaly was the difficulty in computing the distribution of the reference set required for inference. With the advent of computing, Monte Carlo re-randomization takes only seconds, yet the clinical trials culture of invoking a population model has not changed. The second reason is that, under the correct population model, the results of randomization-based inference and population-based inference are typically similar; but this is certainly not always the case under different randomization procedures, heterogeneity, and model misspecification.

As Kempthorne pointed out in the 1950s, the normal theory test should always be considered an approximation to the randomization test, and not vice versa. Randomization tests preserve type I error rates even under heterogeneity, they can be adapted to virtually any type of primary outcome analysis in clinical trials, to multiple treatments, covariate-adjusted analyses, and to confidence intervals. We describe how to do this and conclude that invoking a population model is no longer necessary or desirable in clinical trials practice.

Neyman-Pearson classification: parametrics and sample size requirement

Lucy Xia
The Hong Kong University of Science and Technology
E-mail: lucyxia@ust.hk

Abstract: The Neyman-Pearson (NP) paradigm in binary classification seeks classifiers that achieve a minimal type II error while enforcing the prioritized type I error controlled under some user-specified level α. This paradigm serves naturally in applications such as severe disease diagnosis and spam detection, where people have clear priorities among the two error types. Recently, Tong, Feng, and Li (2018) proposed a nonparametric order statistics based umbrella algorithm that adapts all scoring-type classification methods (e.g., logistic regression, support vector machines, random forest) to respect the given type I error upper bound α with high probability, without specific distributional assumptions on the features and response. Universal the umbrella algorithm is, it demands an explicit minimum sample size requirement on class 0, which is usually the more scarce class. In this work, we employ the parametric linear discriminant analysis (LDA) model and propose a new parametric thresholding algorithm, which does not need the minimum sample size requirements on class 0 observations and thus is applicable to small sample applications such as rare disease diagnosis. Leveraging both the nonparametric and nonparametric thresholding rules, we propose four LDA based NP classifiers, for both low and high dimensional settings. On the theoretical front, we prove NP oracle inequalities for one proposed classifier. This is the first time such theoretical criteria are established under the parametric model assumption and unbounded feature support. Furthermore, as NP classifiers involve a sample splitting step of class 0 observations, we construct a new adaptive sample splitting scheme that can be applied universally to NP classifiers and this

adaptive strategy enhances the accuracy of these classifiers.

Response-adaptive design for clinical trials with recurrent events data

Siu Hung CHEUNG
Southern University of Science and Technology
E-mail: shcheung@cuhk.edu.hk

Abstract: In long‐term clinical studies, recurrent event data are sometimes collected and used to compare the efficacies of two different treatments. The event reoccurrence rates can be compared using the popular negative binomial model, which allows heterogeneity among patients. It is popular that a balanced design in which equal sample sizes are obtained for both treatments is employed. However, it may be desirable to allocate fewer subjects to be assigned to the less‐effective treatment by using a sequential response‐adaptive treatment allocation procedure. Our proposed treatment allocation schemes have been shown to be able to reduce the number of subjects receiving the inferior treatment while at the same time maintains a test power level that is comparable to that of a balanced design. A clinical trial is redesigned to demonstrate the advantages of using our procedure.

S072: Measuring and testing nonlinear dependence

Test for conditional independence with application to conditional screening

Yeqing Zhou
Tongji University
E-mail: yqzhou1991@hotmail.com

Abstract: Measuring and testing conditional dependence are fundamental problems in statistics. Imposing mild conditions on Rosenblatt transformations (Rosenblatt, 1952), we establish an equivalence between the conditional and unconditional independence, which appears to be entirely irrelevant at the first glance. Such an equivalence allows us to convert the problem of testing conditional independence into the problem of testing unconditional independence. We further adopt the Blum–Kiefer–Rosenblatt correlation (Blum et al., 1961) to develop a test for conditional independence, which is powerful to capture nonlinear dependence and is robust to heavy-tailed errors. We obtain explicit forms for the asymptotic null distribution which involves no unknown tunings, rendering fast and easy implementation of our test for conditional independence. With this conditional independence test, we further propose a conditional screening method for high dimensional data to identify truly important covariates whose effects may vary with exposure variables. We use the false discovery rate to determine the screening cutoff. This screening approach possesses both the sure screening and the ranking consistency properties. We illustrate the finite sample performances through simulation studies and an application to the gene expression microarray dataset.

A New Framework for Distance and Kernel-based Metrics in High Dimensions

Xianyang Zhang
Texas A&M University
E-mail: zhangxiany@stat.tamu.edu

Abstract: The paper presents new metrics to quantify and test for (i) the equality of distributions and (ii) the independence between two high-dimensional random vectors. We show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it only detects the equality of means and the traces of covariance matrices in the high-dimensional setup. We propose a new class of metrics which inherit the desirable properties of the energy distance and maximum mean discrepancy/(generalized) distance covariance and the Hilbert-Schmidt Independence Criterion in the low-dimensional setting and is capable of detecting the homogeneity of/completely characterizing independence between the low-dimensional marginal distributions in the high dimensional setup. We further propose t-tests based on the new metrics to perform high-dimensional two-sample testing/independence testing and study their asymptotic behavior under both high dimension low sample size (HDLSS) and high dimension medium sample size (HDMSS) setups. The computational complexity of the t-tests only grows linearly with the dimension and thus is scalable to very high dimensional data. We demonstrate the superior power behavior of the proposed tests for homogeneity of distributions and independence via both simulated and real datasets.

Ball Covariance

Xueqin Wang
Sun Yat-sen University
E-mail: wangxq88@mail.sysu.edu.cn

Abstract: Technological advances in science and engineering have led to the routine collection of large and complex data objects, where the dependence structure among those objects is often of great interest. Those complex objects (e.g., different brain subcortical structures) often reside in some Banach spaces, and hence their relationship cannot be well characterized by most of the existing measures of dependence such as correlation coefficients developed in Hilbert spaces. To overcome the limitations of the existing measures, we propose Ball Covariance as a generic measure of dependence between two random objects in two possibly different Banach spaces. Our Ball Covariance possesses the following attractive properties: (i) It is nonparametric and model-free, which make the proposed measure robust to model mis-specification; (ii) It is nonnegative and equal to zero if and only if two random objects in two separable Banach spaces are independent; (iii) Empirical Ball Covariance is easy to compute and can be used as a test statistic of independence. We present both theoretical and numerical results to reveal the potential power of the Ball Covariance in detecting dependence. Also importantly, we analyze two real datasets to demonstrate the usefulness of Ball Covariance in the complex dependence detection.

Testing the Linear Mean and Constant Variance Conditions in Sufficient Dimension Reduction

Tingyou Zhou
Zhejiang University of Finance & Economics
E-mail: zhoutingyou@163.com

Abstract: Sufficient dimension reduction (SDR, for short) methods characterise the relationship between the response and the covariates, through a few linear combinations of the covariates. Extensive techniques are developed, among which the inverse regression based methods are perhaps the most appealing in practice because they do not involve multi-dimensional smoothing and are easy to implement. However, these inverse regression based methods require two distributional assumptions on the covariates. In particular, the first-order methods, such as the sliced inverse regression, require the linear conditional mean (LCM) assumption, while the second-order methods, such as the sliced average variance

estimation, additionally require the constant conditional variance (CCV) assumption. We propose to check the validity of the LCM and the CCV conditions through mean independence tests, which are facilitated by the martingale difference divergence (MDD). We suggest a consistent bootstrap procedure to decide the critical value of the test. Monte Carlo simulations as well as an application to the horse mussels dataset are conducted to demonstrate the finite sample performances of our proposal.

S073: New developments in statistical methods and inference

Estimating equation methods for longitudinal studies when drop-outs depend on outcome and uncensored observation process

Xia Cui
Guangzhou University
E-mail: cuixia@gzhu.edu.cn

Abstract: In some longitudinal studies, the response variable may be observed at different times for each individual. The associated observation times are often related to the repeated measures and dependent censoring may occur due to death or exclusion from the study related to the disease process. This paper studies inference for semiparametric regression model of longitudinal measurements when observation times are informative and censoring times are dependent. Given covariates or partial covariates, three marginal models are proposed. First, the response variable is described by a partially linear function of covaraites. Second, a proportional rate model is used to fit the intensity of uncensored observation times process. Third, the dependent censoring times are fitted by a transformation regression model. We herein assume the statistical relationship of the error term in the first model, the increment of the uncensored observation times process and the error term in the third model is fully nonparametric. To maintain the homogeneity of the hypothetical error variables under dependent censoring, we appeal to the device of artificial censoring. Based on this, a centralized observation process is proposed. We then estimate the interested parameter by solving an estimating equation constructed based on this centralizing observation process. The proposed estimator is shown to be consistent and asymptotically normal. Simulation studies demonstrate that the proposed inference procedure performs well in many settings. Application to a bladder cancer treatment study is presented.

Statistical Modeling in Non-invasive prenatal screening

Xiaobo Guo
Department of Statistical Science, School of Mathematics, Sun Yat-Sen University
E-mail: mc03gxb@126.com

Abstract: The landmark discovery of fetal cell-free DNA in maternal peripheral blood in 1997 laid the foundation for non-invasive prenatal screening (NIPS). In practice, Z-score is the prevailing method used in NIPS. However, the strict yet clear statistical basis for the NIPS area is still limit. In this talk, I will discuss how to develop the statistical modeling framework for the NIPS， and develop a novel noninvasive prenatal diagnosis method. Under the proposed modeling framework, we provide theoretical answer to a series of questions arisen in NIPS. We also use simulation and real data analysis to show that our proposed method can substantially reduce the false negative rate for the samples with low fetal fraction.

Model Confidence Bounds for Variable Selection

Yang Li
Renmin University of China
E-mail: yang.li@ruc.edu.cn

Abstract: In this article, we introduce the concept of model confidence bounds (MCBs) for variable selection in the context of nested models. Similarly to the endpoints in the familiar confidence interval for parameter estimation, the MCBs identify two nested models (upper and lower confidence bound models) containing the true model at a given level of confidence. Instead of trusting a single selected model obtained from a given model selection method, the MCBs proposes a group of nested models as candidates and the MCBs' width and composition enable the practitioner to assess the overall model selection uncertainty. A new graphical tool — the model uncertainty curve (MUC) — is introduced to visualize the variability of model selection and to compare different model selection procedures. The MCBs methodology is implemented by a fast bootstrap algorithm that is shown to yield the correct asymptotic coverage under rather general conditions. Our Monte Carlo simulations and a real data example confirm the validity and illustrate the advantages of the proposed method.

Testing of covariate effects under ridge regression for high-dimensional data

Xu Liu
Shanghai University of Finance and Economics
E-mail: liu.xu@sufe.edu.cn

Abstract: In this paper, we revisit the ridge estimation in high-dimensional regression models. We propose a novel estimator of the error's variance and give its asymptotic result based on the random matrix theories as the dimension of covariates diverges with the sample size. {This estimator is promising compared with its competitors including the refitted cross validation method in many scenarios. Meanwhile, an upper bound of mean squared error is given for the ridge estimator of regression coefficients, and an efficient method is proposed on testing the high-dimensional covariate effects.The proposal is valid under both low- and high-dimensional models, which performs well not only for the sparse alternatives but also for the non-sparse ones. Numerical examples are used to assess the finite-sample performance of the proposed method.

Bayesian Variable Selection for Linear Regression with Interaction Terms

Wensheng Zhu
Northeast Normal University
E-mail: wszhu@nenu.edu.cn

Abstract: In modern statistics, we always encounter high-throughput data with huge number of covariates or features for a small number of subjects. In many cases, researchers show a great interest in the interactions of these covariates. However, it is a challenging task to construct the suitable model by selecting the active covariates among tens of thousands of interactions. In this talk, we propose a Bayesian variable selection approach to identify interactions in the presence of huge dimensional covariates. Spike and slab Gaussian priors are used on the main effects as well as interactions, which shrink and diffuse, respectively as the sample size increases. To reduce computational complexity, our method can be carried out by the Gibbs sampler called "Skinny Gibbs", which was proposed recently in JASA. We show that Skinny Gibbs sampler has a stationary distribution and also exhibits the strong model selection consistency even when p=exp(o(n)).

Simulations in genetic association studies indicate that our proposed method offers merits in the detection of gene-gene and gene-environmental interactions.

S074: New Modeling Methods for Time Series of Analysis

Flexible bivariate Poisson integer-valued GARCH model

Fukang Zhu
Jilin University
E-mail: zfk8010@163.com

Abstract: Integer-valued time series models have been widely used, especially integer valued autoregressive (INAR) models and integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models. Recently, there has been a growing interest in multivariate count time series. However, existing models restrict the dependence structures imposed by the way they constructed. In this paper, we consider a class of flexible bivariate Poisson INGARCH(1,1) model whose dependence is established by a special multiplicative factor. Stationarity and ergodicity of the process are discussed. The maximization by parts algorithm and its modified version together with the alternative method by using R package Template Model Builder are employed to estimate the parameters of interest. The consistency and asymptotic normality for estimates are obtained and the finite sample performance of estimators are given via simulations. A real data example is also provided to illustrate the model.

Detecting mean increases in zero truncated INAR(1) processes

Cong Li
Jilin University
E-mail: li_cong@jlu.edu.cn

Abstract: Count data with zero truncation are common in the production process. It's essential to monitor these data during production flow, production quality control and market management. Most of the previous studies were based on the independent observations assumption. In fact, serial dependence of count data which significantly affects the performance of the control charts exists extensively in practice. Motivated by this, several important first-order integer-valued autoregressive time series processes are used to model the autocorrelated count data with zero truncation. We investigate the effectiveness of three following charts, the combined jumps chart, the exponentially weighted moving average chart and the cumulative sum chart, to detect the upward shifts of the process mean based on these models. A bivariate Markov chain approach could be used to obtain the average run length of these charts. Design recommendations for achieving robustness are provided based on the computation study. An application to product quality complaints data is presented to demonstrate good performances of the charts.

Random coefficients self-exciting threshold integer-valued autoregressive processes driven by logistic regression

Kai Yang
Changchun University of Technology
E-mail: yangkai@ccut.edu.cn

Abstract: In this article, we introduce a new first-order random coefficients self-exciting threshold integer-valued autoregressive processes. The autoregressive coefficients are driven by a logistic regression structure, so that the explanatory variables can be included. Basic probabilistic and statistical properties of this model are discussed. Conditional least squares and conditional maximum likelihood estimators, as well as the asymptotic properties of the estimators are discussed. The nonlinearity test problem is also addressed. As an illustration, we evaluate our estimates through a simulation study. Finally, we apply our method to the data sets of sexual offences in Ballina, New South Wales (NSW), Australia, with two covariates of drug offences and temperature. The result reveals that the proposed model fits the data sets well.

Threshold negative binomial autoregressive model

Mengya Liu
Jilin University
E-mail: lmy1768470589@163.com

Abstract: This article studies an observation-driven model for time series of counts, which allows for overdispersion and negative serial dependence in the observations. The observations are supposed to follow a negative binomial distribution conditioned on past information with the form of thresh old models, which generates a two-regime structure on the basis of the magnitude of the lagged observations. We use the weak dependence approach to establish the stationarity and ergodicity, and the inference for regression parameters are obtained by the quasi-likelihood. Moreover, asymptotic properties of both quasi-maximum likelihood estimators and the threshold estimator are established, respectively. Simulation studies are considered and so are two applications, one of which is the trading volume of a stock and another is the number of major earthquakes.

S075: Massive regression analysis

Additive partially linear models for massive heterogeneous data

Binhuan Wang
NYU School of Medicine
E-mail: binhuan.wang@nyumc.org

Abstract: We consider an additive partially linear framework for modelling massive heterogeneous data. The major goal is to extract multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. We propose an aggregation type of estimators for the commonality parameters that possess the asymptotic optimal bounds and the asymptotic distributions as if there were no heterogeneity. This oracle
result holds when the number of sub-populations does not grow too fast and the tuning parameters are selected carefully. A plugin estimator for the heterogeneity parameter is further
constructed, and shown to possess the asymptotic distribution as if the commonality information were available. Furthermore, we develop a heterogeneity test for the linear components and a homogeneity test for the non-linear components accordingly. The performance of the proposed
methods is evaluated via simulation studies and an application to the Medicare Provider Utilization and Payment data. Keywords and phrases: Divide-and-conquer, homogeneity, heterogeneity, oracle property, regression splines.

Discrepancy between global and local principal component analysis on large-panel high-frequency data

Xinbing Kong
Nanjing Audit University
E-mail: xinbingkong@126.com

Abstract: The global principal component analysis (GPCA), PCA applied to the whole sample, is not reliable to reconstruct the common components of a large-panel high-frequency data when the factor space is time-varying, but it works when the factor space does not change in the time domain. The local principal component analysis (LPCA), PCA carried on subsamples,

results in consistent estimates of the common components even if the factor loading processes follow continuous-time It^{o} semimartingales, but it loses efficiency when the factor space is time invariant. This motivates us to study the discrepancy between the GPCA and LPCA in recovering the common components of a large-panel high-frequency data. In this paper, we measure the discrepancy by the total sum of squared differences between common components reconstructed from GPCA and LPCA. The asymptotic distribution of the discrepancy measure is provided when the factor space is time invariant and the dimension p and the sample size n tends to infinity simultaneously. Alternatively when some factor loadings are time-varying, the discrepancy measure explodes in a rate higher than $sqrt{pk^{3/2}_n/n}$ under some mild signal conditions on the magnitude of time-variation of the factor loadings, where k_n is the size of each subsample. We apply the theory to test the hypothesis that the factor space does not change in time. We show that the test performs well in controlling the type I error and detecting time-varying factor spaces. This is checked by extensive simulation studies. A real data analysis provides strong evidence that the factor space is always time-varying within a time span longer than one week.

Spatial-Temporal Prediction of PM2.5 concentration in North China Plain Using the Machine Learning

Bin Guo
Southwestern University of Finance and Economics
E-mail: 470695010@qq.com

Abstract: In this study, we used the outputs of Community Multi-scale Air Quality Modeling (CMAQ) system, Aerosol Optical Depth (AOD) and the weather information to forecast the PM2.5 concentration in North China Plain. We compare the Linear Regression, SVM, Random Forest et al. machine learning methods in this study. Some deep learning algorithms are also implemented. The machine learning methods show promising performances.

Identifying sensitive subset based on ultra-high dimensional correlated covariates for survival data

Ye He
University of Electronic Science and Technology of China
E-mail: hey0624@126.com

Abstract: In this paper, we consider a single-index threshold Cox proportional hazard model to identify patients who are sensitive to a specific treatment, based on ultra-high dimensional correlated covariates.
A smoothed partial likelihood is proposed to estimate the parameters, and furthermore a functional regularization technique is used to select the effective region of a treatment. Asymptotic analysis reveals that the proposed estimator is consistent and can identify null subregions with probability tending to one. The proposed approach is evaluated through two simulation studies and an application to analyze TCGA skin melanoma data.

S076: Special Invited Session -DiDi Session- Statistical Challenges and Opportunities in Ride-sharing Platform

Challenges in Analyzing Two-sided Market and Its Application on Ride-sourcing Platform

Hongtu Zhu
UNC Chapel Hill
E-mail: zhuhongtu@didiglobal.com

Abstract: In this talk, we will introduce a general analytical framework for large scale data obtained from two-sided markets, especially ride-sourcing platforms like DiDi. This framework integrates classical methods including Experiment Design, Causal Inference and Reinforcement Learning, with modern machine learning methods, such as Graph Convolutional Models, Deep Learning, Transfer Learning and Generative Adversarial Network. We aim to develop fast and efficient approaches to address five major challenges for ride-sharing platform, ranging from demand-supply forecasting, demand-supply diagnosis, MDP-based policy optimization, A-B testing, to business operation simulation. Each challenge requires substantial methodological developments and inspires many researchers from both industry and academia to participate in this endeavor. Based on our preliminary results for the policy optimization challenge, we received the Daniel Wagner Prize for Excellent in Operations Research Practice in 2019. All the research accomplishments presented in this talk are joint work by a group of researchers at Didi Chuxing and our international collaborators.

Marry Data and AI with SQLFlow

Ziyao Gao
Senior Data Scientist at DiDi
E-mail: gaoziyao@didiglobal.com

Abstract: SQLFlow is an open-source machine learning tool co-developed by Ant Financial and DiDi, aiming to bridge the SQL engines and AI engines like TensorFlow. Since introduced, SQLFlow has dramatically lowered the bar for DiDi's operations/business analysts to leverage advanced ML capabilities and assisted them to do data collection, model training and predicting, as well as model explanation in one stop. Today's presentation will start with a brief introduction of SQLFlow, followed by three real-world applications including how SQLFlow assists DiDi's operation analysts to identify target riders for marketing campaign, to uncover most decisive factors for user retention, and to cluster drivers into different segments based on their daily riding patterns. In the near future, as a wide body of ML models are added and system integration is improved, SQLFlow will pave the way for DiDi to transit from data-driven to data-intelligence driven.

Driving Risk Assessment for Ride-hailing Drivers

Liang Shi
Virginia Tech Transportation Institute
E-mail: sliang@vt.edu

Abstract: The rise of ride-hailing services in the last decade has opened a new mode of travel and provided work opportunities for millions of drivers. Our study evaluates crash risk factors associated with ride-hailing drivers, including crash history and ride-hailing operational characteristics. We utilize the Poisson Generalized Additive Model to accommodate the potential nonlinear relationship between the logarithm of crash rate and risk factors. Results show that crash history, the percentage of long-shift orders, driving distance, operations during peak hours, years of being a ride-hailing driver, and passenger rating are significantly associated with crash rate. Among them, several factors display nonlinear relationship with the logarithm of crash rate. The SHapley Additive exPlanation method is used to explain and visualize the contribution of each risk fa

ctor. The results indicate that crash history, years of being a ride-hailing driver, and total driving distance are the leading f actors contributing to ride-hailing driver crash risk. The results of this study provide valuable information for understanding crash risk for ride-hailing drivers and for developing safety co untermeasure and ride-hailing driver education programs.

S077: Robust and efficient modern regression in high dimensional and complex data

Total Variation Regularized Frechet Regression for Metric-Space Valued Data

Zhenhua Lin
National University of Singapore
E-mail: stalz@nus.edu.sg

Abstract: Non-Euclidean data that are indexed with a scalar predictor such as time are increasingly encountered in data applications, while statistical methodology and theory for such random objects are not well developed yet. To address the need for new methodology in this area, we develop a total variation regularization technique for nonparametric Frechet regression, which refers to a regression setting where a response residing in a generic metric space is paired with a scalar predictor and the target is a conditional Frechet mean. Specifically, we seek to approximate an unknown metric-space valued function by an estimator that minimizes the Frechet version of least squares and at the same time has small total variation, appropriately defined for metric-space valued objects. We show that the resulting estimator is representable by a piece-wise constant function and establish the minimax convergence rate of the proposed estimator for metric data objects that reside in Hadamard spaces. We illustrate the numerical performance of the proposed method for both simulated and real data, including the metric spaces of symmetric positive-definite matrices, probability distributions and phylogenetic trees.

Bayesian Covariate-dependent Gaussian Graphical Model

Yingying Wei
The Chinese University of Hong Kong
G-mail: yweicuhk@gmail.com

Abstract: There has been active research on estimating a single Gaussian graphical model for a set of samples. However, when there exists heterogeneity among the samples, learning a single graphical model for all of the samples can lead to many spurious edges. The recent emerging methods on joint modeling of multiple graphical models allow graphical structures to change with univariate, categorical covariates. However, they cannot handle continuous covariates, let alone multiple covariates. The early proposed frequentist approach to linking multiple covariates to graphical structures must first partition the space of covariates and then separately learn graphs on each portion of the data, resulting in unstable estimators and a loss of interpretability for the covariates.

Here, we propose a novel Bayesian framework to study how the graphical structures change with covariates for Gaussian graphical models. Our proposed method can handle all types of covariates, borrow strength across the whole covariate space to improve edge detection, and provide direct interpretation for effects of covariates on each edge of the graph. We develop an efficient parallel Markov chain Monte Carlo algorithm to conduct posterior inference. We applied the proposed method to study how gene regulatory networks vary across different types of covariates such as age and disease categories.

Pairwise-rank-likelihood methods for the semiparametric transformation model

Tao Yu
National University of Singapore
E-mail: stayt@nus.edu.sg

Abstract: In this paper, we study the linear transformation model in the most general setup. This model includes many important and popular models in statistics and econometrics as special cases. Although it has been studied for many years, the methods in the literature either are based on kernel-smoothing techniques or make use of only the ranks of the responses in the estimation of the parametric components. The former approach needs a tuning parameter, which is not easily optimally specified in practice; and the latter approach may be {less accurate and computationally expensive}. In this paper, we propose a {pairwise rank likelihood} method {and extend it to a score-function-based method. Our methods estimate} all the unknown parameters in the linear transformation model, and we {explore the theoretical properties of} our proposed estimators. Via extensive numerical studies, we demonstrate that {our methods are} appealing in that the estimators are not only robust to the distribution of the random errors but also {in many cases more accurate} than those of the existing methods.

Predicting time series with abrupt changes and smooth evolutions

Jie Ding
University of Minnesota
E-mail: dingj@umn.edu

Abstract: A methodology (referred to as kinetic prediction) is introduced for predicting time series undergoing unknown abrupt changes or smooth evolutions in their data generating distributions. Based on Kolmogorov epsilon-entropy, we propose a concept called epsilon-predictability that quantifies the size of a model class and the maximal number of structural changes that guarantee the achievability of asymptotically optimal prediction. Moreover, for parametric distribution families, the aforementioned kinetic prediction with discretized function spaces is extended to its counterpart with continuous function spaces, which naturally leads to an efficient sequential Monte Carlo implementation. Wide applicability of the proposed methodology will be illustrated by its applications to time-varying cointegration, time-varying volatility models, and case studies in finance.

S078: Joint Modeling and Classification Models for Complex Biomedical Data

Multilevel joint modeling of hospitalization and survival in patients on dialysis

Esra Kurum
University of California
E-mail: esra.kurum@ucr.edu

Abstract: More than 720,000 patients with end-stage renal disease in the U.S. require life-sustaining dialysis treatment that is predominantly received at local dialysis facilities. In this population of typically older patients with a high morbidity burden, hospitalization is frequent at a rate of about twice per patient-year. Aside from frequent hospitalizations, which is a major source of mortality, overall mortality in dialysis patients is higher than other comparable populations, including Medicare patients with cancer. Thus, understanding patient- and facility-level risk factors that jointly contribute to longitudinal hospitalizations and mortality is of interest. Towards this

objective, we propose a novel methodology to jointly model hospitalization, a binary longitudinal outcome, and survival, based on multilevel data from the United States Renal Data System (USRDS), with repeated observations over time nested in patients and patients nested in dialysis facilities. In our approach, the outcomes are modeled through a common set of multilevel random effects. In order to accommodate the USRDS data structure, we depart from the literature on joint modeling of longitudinal and survival data by including multi-level random effects and multilevel covariates, at both the patient and facility levels. An approximate EM algorithm is developed for estimation where fully exponential Laplace approximations are utilized to address computational challenges. Standard error formulas for the estimated parameters are derived and evaluated to guide practical inference.

Construction, Visualization and Application of Neutral Zone Classifiers

Daniel Jeske
University of California
E-mail: daniel.jeske@ucr.edu

Abstract: When the potential for making accurate classifications with a statistical classifier is limited, a neutral zone classifier can be constructed by adding a no-decision option as a classification outcome. We show how a neutral zone classifier can be constructed from a receiving operating characteristic (ROC) curve. We extend the ROC curve graphic to highlight important performance characteristics of a neutral zone classifier. Additional utility of neutral zone classifiers is illustrated by showing how they can be incorporated into the first stage of a two-stage classification process. At the first stage, a classification is attempted from easily collected or inexpensive features. If the classification falls into the neutral zone, additional relatively more expensive features can be obtained and used to make a definitive classification at the second stage. The methods discussed in the paper are illustrated with an application pertaining to prostate cancer.

Misspecification of a dependent variable in the logistic model controlling for the repeated longitudinal measures

Yi-Ting Hwang
National Taipei University
E-mail: hwangyt@gm.ntpu.edu.tw

Abstract: Many medical applications are interested to know the disease status which is often related to multiple serial measurements. Precise measurements for the binary outcome are required when using the maximum likelihood estimation. To incorporate all the data information in the estimation, Hwang et al. (2015) derived the joint likelihood of the observed data. Nevertheless, the binary data might be often mismeasured owing to various reasons. When binary outcomes are subject to misclassification, the estimators of the coefficients of the logistic regression are biased. To reduce the bias, this paper incorporates the misspecification in the joint likelihood function. The joint likelihood approach along with the EM algorithm is used to find the estimates. Monte Carlo simulations are conducted to compare the impact of misspecification on the estimates. A retrospective data for the recurrence of AF is used to illustrate the usage of the proposed model.

Joint Trajectories with Variable Selection

Wendy Lou
University of Toronto
E-mail: wendy.lou@utoronto.ca

Abstract: Patterns of change over time can serve as the basis for identifying subtypes of diseases in a heterogeneous population, and utilizing data from multiple sources allows for the simultaneous study of longitudinal patterns. Motivated by a birth cohort study, we will discuss approaches based on mixture models and machine learning, along with variable-selection strategies, to classify children in terms of biological risks, clinical pathways and environmental exposures for developing chronic conditions, such as asthma. To illustrate the methodology, numerical results from both real data and simulations will be presented. (This is a joint work with Zihang Lu).

S079: Statistical Advancements for Emerging Challenges in Health Data Science

Identifying the Best Predictive SNP in GWAS for Companion Diagnostics

Xinping Cui
Department of Statistics, University of California
E-mail: xinping.cui@ucr.edu

Abstract: It is now well recognized that the effectiveness and potential risk of a treatment often vary by patient subgroups. A companion diagnostic (CDx) is a diagnostic test co-developed with drugs for the drug safety and effectiveness. Most FDA-approved CDx tests aim for identifying a single, specific predictive SNP, and building ``one drug, one SNP'' model. Predictive biomarkers are of particular importance for clinicians to select right treatment at the right dose for the right person at the right time for the right outcome. Typically, SNPs are ranked in terms of their p-values, and an easy and intuitive way is to select the "best" SNP with the smallest p-value. However, the p-value ordering is sensitive to the noise, and doesn't necessarily correlate with the true ordering of SNPs. Furthermore, to our best knowledge, no work has considered the response profiles of the SNPs. The SNP, even with the smallest p-value, would not be useful for a CDx unless its response profile satisfies certain patterns. In this paper, we first develop a score function to quantify the predictive ability of each candidate SNP. We then formulate the parameters of interest as the minimal difference in predictive ability between a candidate SNP and the best SNP conditional on SNP response profiles satisfying desired patterns. We propose to find the ``best'' SNP based on simultaneous confidence set of parameters of interest built upon the framework of multiple comparison with the best (MCB) controlling per family error rate (PFER). Simulation studies and application in Alzheimer Disease randomized clinical trials will be used to demonstrate the novel discovery and advantage of the proposed framework.

Statistical approaches for identifying biomarkers for a group of cancer drugs

Jian Zhang
University of Kent
E-mail: J.Zhang-79@kent.ac.uk

Abstract: We propose a novel approach to nonparametric variable screening for sparse multivariate additive models with random effects, which includes two stages. In Stage 1, each nonparametric component is approximated by a linear combination of spline basis functions. Under this approximation, the above screening problem can be treated as selecting block-matrices of regression coefficients for a multivariate regression model. In Stage 2, a series of filtering operations are conducted by projections of the multiple response observations into the covariate space;

each filter is tailored to a particular covariate and resistant to interferences originating from other covariates and from background noises. The filtering is further improved by sequentially nulling significant covariates detected in the previous steps. An asymptotic theory on the selection consistency has been established under some regularity conditions. By simulations, the proposed procedure is shown to outperform the existing procedures in terms of sensitivity and specificity over a wide range of scenarios. We apply the proposed approach to the integrative analysis of the anti-cancer drug data, identifying a few biomarkers that potentially influence the concentration of drugs in cancer cell lines.

Computational methods to elucidate chromatin topological structures using 3D genomic maps

Shihua Zhang
Academy of Mathematics and Systems Science, CAS
E-mail: zsh@amss.ac.cn

Abstract: T The chromosome conformation capture (3C) technique and its variants have been employed to reveal the existence of a hierarchy of structures in three-dimensional (3D) chromosomal architecture, including compartments, topologically associating domains (TADs), sub-TADs and chromatin loops. In this talk, I am going to introduce three methods on deciphering 3D genomic maps: (1) a mixed-scale dense convolutional neural network model (HiCMSD) to enhance low-resolution Hi-C interaction map for deciphering accurate multi-scale topological structures; (2) a generic and efficient method to identify multi-scale topological domains (MSTD), including cis- and trans-interacting regions, from a variety of 3D genomic datasets; (3) a powerful and robust circular trajectory reconstruction tool CIRCLET without specifying a starting cell for resolving cell cycle phases of single cells by considering multi-scale features of chromosomal architectures.

[1] Ye Y, Gao L, **Zhang S**. MSTD: an efficient method for detecting multi-scale topological domains from symmetric and asymmetric 3D genomic maps. **Nucleic Acids Research** (2019), 47: e65.

[2] Ye Y, Gao L, **Zhang S**. Circular trajectory reconstruction uncovers cell-cycle progression and regulatory dynamics from single-cell Hi-C maps. **Advanced Science** (2019), doi:10.1002/advs.201900986

Deep learning for decoding molecular phenotypes with radiogenomics in breast cancer

Pingzhao Hu
University of Manitoba
E-mail: pingzhao.hu@umanitoba.ca

Abstract: Objective

It has been believed that traditional handcrafted radiomic features extracted from magnetic resonance imaging (MRI) of tumors are normally shallow and low-ordered. Recent advancement in deep learning technology shows that the high-order deep radiomic features extracted automatically from tumor images can capture tumor heterogeneity in a more efficient way. We hypothesize that MRI-based deep radiomic phenotypes have significant associations with molecular profiles of breast cancer tumors. We aim to identify MRI-based signatures that can explain the potential underlying genetic mechanisms and predict the molecular classification of invasive breast cancers.

Methods

We develop a new deep learning model to retrospectively extract 4,096 MRI-based radiomic phenotypes from breast cancer tumors collected by The Cancer Imaging Archive (TCIA). These phenotypes of tumors are associated with genomic features (commercialized gene signatures, expression of risk genes, and pathways activities) of the corresponding molecular profiles (e.g. gene expression) and other clinical features collected from The Cancer Genome Atlas (TCGA). We develop novel association and classification methods to select the most-predictive radiogenomic features for the clinical phenotypes, including tumor size (T), lymph node metastasis(N) from breast cancer TNM staging system which is wildly used in clinic, and status of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2).

Results

We find that transcriptional activities of various genetic pathways and gene signatures are positively associated with more than 1000 of the 4,096 MRI-based radiomic phenotypes. These radiomic phenotypes are also associated with the mRNA expression of the risk genes identified from other two genome-wide association studies. Higher performances are obtained in the prediction of HER2 status, ER status and tumor size(T) than PR status and lymph node metastasis(N). These identified MRI-based radiomic phenotypes also show significant power to stratify the breast cancer tumors, which may have a significant clinical impact.

Conclusion

Our radiogenomic approach for identifying MRI-based imaging signatures may pave potential pathways for the discovery of genetic mechanisms regulating specific tumor phenotypes and may enable a more rapid innovation of novel imaging modalities, hence accelerating their translation to personalized medicine.

S080: Matrix Estimation and Matrix regression

Estimation of error variance via ridge regression

Xingdong Feng
Shanghai University of Finance and Economics, China
E-mail: feng.xingdong@mail.shufe.edu.cn

Abstract: We propose a novel estimator of error variance and establish its asymptotic properties based on ridge regression and random matrix theory. The proposal is valid under both low- and high-dimensional models, and performs well not only for non-sparse cases but also for sparse ones. We assess the finite-sample performance of the proposed method in an intensive numerical study, which indicates that it is promising compared with its competitors in many interesting scenarios.

Nonparametric Regression with a Randomly Censored Independent Variable

Lei Huang
Southwest Jiaotong University
E-mail: stahl@swjtu.edu.cn

Abstract: Censoring occurs often in data collection. In this paper, we consider the nonparametric regression when the covariate is censored. In contrast to censoring in the response variable as does in most survival analysis, regression with censored covariates is more challenging. We propose to estimate the regression function using conditional hazard rates. Asymptotic normality of our proposed estimator is established. Compared with that based on complete cases, both theoretical results and simulation studies demonstrate that the proposed nonparametric method could estimate unknown regression functions more efficiently especially with high censoring rate. We illustrate and compare methods using the well-known dataset from a randomized placebo controlled clinical trial of the drug

D-penicillamine (DPCA).

Influence Matrix Analysis

Wei Lan
School of Statistics, Southwest University of Finance and Economic
E-mail: lanwei@swufe.edu.cn

Abstract: This paper introduces the influence matrix regression model (IMR), which establishes the relationship between the influence matrix of actors and a set of similarity matrices induced by their associated attributes. This model not only extends the commonly used spatial autoregressive model to incorporate the influence matrix, but also allows the influence matrix to change with time. We then employ the quasi-maximum likelihood estimation method to estimate unknown regression coe cients. The resulting estimator is asymptotically normal without imposing the normality assumption. When the number of similarity matrices is large, a BIC-type criterion is employed to select relevant matrices. To assess the adequacy of the proposed model, we further propose an influence matrix test, and develop a novel approach to obtain the limiting distribution of the test. The simulation studies support our theoretical findings, and an empirical example is presented to illustrate the usefulness of the proposed IMR model.

Supervised cluster analysis of non-Gaussian functional data

Jiakun Jiang
Tsinghua University
E-mail: jiakunj@tsinghua.edu.cn

Abstract: In this paper we study cluster analysis of functional regression with a ran-dom response curve and vector covariates. We propose a mixed transformation functional regression model with an unknown number of clusters. Compared to the existing cluster analysis of functional regression, our model has several advantages. First, our model is free from normality assumption. Second, it is supervised, in that the clustering is based on the relationship between the functional response and covariates. Finally, we allow the number of clusters to be unknown a priori. We propose a combination of penalized likelihood and estimating equation methods to estimate the number of clusters, regression parameters and transformation function simultaneously. We establish theoret-ical properties, including sqrt(n) consistency and asymptotic normality, for the proposed estimators. Extensive simulation results show that the proposed esti-mation procedure works very well. The proposed method is utilized to analyze housing market conditions in China from 2007 to 2014, which leads to some interesting ndings.

S081: Design and Analysis for Medical Studies with Practical Illustrations

Robust Design Approaches in Biomedical Research

Timothy O'Brien
Loyola University Chicago
E-mail: tobrie1@luc.edu

Abstract: Researchers often find that nonlinear regression models are more applicable for modelling various biomedical phenomena than are linear ones since they tend to fit the data well and since these models (and model parameters) are more scientifically meaningful. For example, researchers in fields as diverse as toxicology, pharmacology, biometry, and medicine typically fit four-parameter sigmoidal functions and are often in a position of requiring optimal or near-optimal designs for the chosen nonlinear model. A common shortcoming of most optimal designs for nonlinear models used in practical settings, however, is that these designs typically focus only on (first-order) parameter variance or predicted variance, and thus ignore the inherent nonlinear of the assumed model function. Another shortcoming of optimal designs is that they often have only p support points (where p is the number of model parameters), and so cannot be used to test for model adequacy.

This talk reviews and underscores the practical advantages of (generalized and normal-based) nonlinear models, and examines various robust design criteria, including geometric and uniform design strategies given in O'Brien et al (2009) and O'Brien (2016), reflection designs in O'Brien and Silcox (2019), as well as those based on second-order (curvature) considerations. Several key examples are provided to illustrate these ideas using commonly-used software such as SAS and R.

Analyzing longitudinal activity data collected from wearable devices

Meike Niederhausen
Oregon Health & Science University
E-mail: niederha@ohsu.edu

Abstract: Fitness trackers have become a popular device for researching associations between physical activity and health status. We will present challenges in curating and analyzing data from wearable devices. Data visualizations were key in this process to detect data anomalies, plot activity trends, and understand complex relationships amongst variables. Since many activity studies only collect data 2-7 days, one question in particular we investigated is how long to track free-living individuals to derive reliable estimates of their usual physical activity levels. This was done using changepoint detection techniques on long data streams of activity data. The data are from a 6-month workplace study with 431 healthy volunteers that collected minute-by-minute activity from a wearable device along with biometric and cardiometabolic measures at the start and end of study. Tracker data included heart rate, steps, body states (sleep, inactive, light activity, moderate activity, and off wrist), and sleep metrics.

A case study of refining testing strategy using graphical approach

Eva Hua
Novartis
E-mail: eva.hua@novartis.com

Abstract: For clinical trials with multiple treatment arms or endpoints a variety of sequentially rejective, weighted Bonferroni-type tests have been proposed, such as gatekeeping procedures, fixed sequence tests, and fallback procedures. The graphic approach (Bretz et al., 2009) is a simple iterative method to construct and perform such Bonferroni-type tests, where the resulting multiple test procedures are represented by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. We apply the graphic approach to a complex case study with multiple endpoints, multiple doses and multiple comparators to tailor the multiple test procedure to given study objectives. It is a good communication tool for the team with clinicians and medical doctors on this complicated multiple testing strategy. The hierarchy of the decision points with the alpha splitting of the hypotheses is

described for each of the potential options. The overall power calculation based on the multiple testing strategy takes the correlations between hypotheses for the same endpoints into consideration, as well as the correlations between different endpoints.

Bayesian single-index joint models of multivariate longitudinal and survival data

An-Min Tang
Department of Statistics, Yunnan University
E-mail: tam13as@sina.com

Abstract: We propose a joint model for multivariate longitudinal and survival data. One main feature of the posited model is that we relax the commonly used linear or nonlinear assumption for trajectory functions shared by longitudinal processes and survival processes, by using partial linear single-index model to specify these functions. Based on our proposed feasible high-efficient algorithm for computing survival and penalized splines for link function in single-index model, a Bayesian approach is proposed to simultaneously obtain Bayesian estimates of unknown parameters, random effects and non-parametric functions by combining the Gibbs sampler and the Metropolis-Hastings algorithm. Several simulation studies and an example are presented to illustrate the propose methodologies. studies and a real example are used to illustrate our proposed methodologies.

S082: Complex Innovative Designs in Practice of Early Phase Drug Development

Model-based Phase I Designs with Incorporation of Individualized Dosing Using Toxicity Scores from Multiple Treatment Cycles

Jun Yin
Mayo Clinic
E-mail: yin.jun@mayo.edu

Abstract: In the era of molecularly targeted agents (MTAs) and immunotherapies, several aspects of phase I designs need to evolve in order to adapt to the changing nature of cancer therapies and to expedite their clinical translation. We have previously developed novel phase I designs to incorporate evaluation of early efficacy, in addition to clinician-assessed toxicity scores from multiple treatment cycles, for identification of tolerable and efficacious doses for subsequent investigation. To extend our previous work, we introduced a dose algorithm that allows patients in subsequent treatment cycles to be treated with an individualized dose that is tailored to their specific tolerance and the cumulative adverse effects of the drug. The proposed method provides a comprehensive statistical framework for the personalized dose modification in subsequent treatment cycles. The design is calibrated with respect to specific operating characteristics. We conduct extensive simulations to assess the performance of the proposed design with comparison to our previous published work.

A New Bayesian Framework for Master Protocols with Type I Error Control

Yuan Ji
The University of Chicago
E-mail: yji@health.bsd.uchicago.edu

Abstract: We consider a new Bayesian framework for master protocols in which hypotheses are treated as random quantities under a Bayesian testing setting. Several benefits come with this framework. First, sample size, operating characteristics, and frequentist properties can be assessed via simulation. Second, Bayesian properties such as conditional probabilities of making a wrong decision, or Bayesian type I error rates can be evaluated. Lastly, Bayesian multiplicity control can be applied based on prior calibration. We show the benefits in terms of frequentist property, e.g., sample size saving, and also in terms of Bayesian property, e.g., the use of prior to control probability of making errors. Examples of real-world trials will be provided. The construction of the Bayesian models involve application of parametric and nonparametric priors. The models can accommodate different types of trials, such as phase 1b expansion cohorts, basket and umbrella trials, and any multi-arm trials with or without the same endpoints.

S083: Novel Complex Data Analysis Methods

Regionalization of PM2.5 in Jing-Jin-Ji Area Using Convex Clustering

Hui Huang
Sun Yat-sen University
E-mail: huangh89@mail.sysu.edu.cn

Abstract: For air-pollution control, it is important to specify regions with similar emission patterns so that more precise local policies can be made accordingly. In this study, we treat PM2.5 concentrations from monitoring stations and/or computer model grid cells as spatially-dependent functional data. Geographical information of station/grid locations is used to define a graph, and we develop a modified convex clustering method to group these locations. In numerical studies, we find that the conventional ADMM is fast enough to deal with large spatiotemporal datasets. The proposed method is applied to regionalize PM2.5 concentrations in Jing-Jin-Ji, one of the most polluted area in China. Results show that there are three different emission patterns in this area with clear boundaries.

Genome-wide Association Testing for Pleiotropic Effects using GWAS Summary Statistics

Zhonghua Liu
The University of Hong Kong
E-mail: zhhliu@hku.hk

Abstract: It has been empirically observed that many human complex traits are genetically correlated, indicating that a single gene might affect multiple phenotypes. This biological phenomenon is referred to as “pleiotropy”, which is of increasing scientific and medical interests for precision medicine and drug repurposing. However, there are very few statistical methods available for detecting such pleiotropic genetic variants. We note that the null hypothesis of no pleiotropic effect is composite and thus imposes great statistical challenges. In this paper, we develop a novel pleiotropic effect test (PET) to detect SNPs with pleiotropic effects on two phenotypes by taking the composite null issue into account. Extensive simulation studies show that our method maintains correct type I error rate and is well powered. We also apply the PET to the global lipids GWAS data sets and detect many novel pleiotropic SNPs affecting lipids traits. We also develop an user-friendly and computationally efficient R package PET for public use.

Selection models for the efficient design of family studies

Yujie Zhong
School of Statistics and Management, Shanghai University of Finance and Economics
E-mail: zhong.yujie@mail.shufe.edu.cn

Abstract: Family studies on the genetic basis for disease typically recruit a random sample individuals from a disease registry with their respective relatives to obtain genetic and phenotypic data. Individuals in the disease registry, however, may be viewed as comprising a phase I sample in a two-phase study. We propose the use of selection models to exploit this phase I data and more efficiently identify a phase II sample of families for study. Copula models are adopted characterize the within family dependence in disease onset times but the likelihood is constructed based on current status data in order to rely on the physician diagnosis when family members are examined. Empirical studies are conducted to demonstrate the efficiency gains that can be realized over simple random or balanced sampling schemes and to study the effect of misspecifying the design parameters.

Semiparametric Varying-coefficient Study of Mean Residual Life Models with right-censored and length-biased data

Fangfang Bai
University of International Business and Economics
E-mail: bff03021@163.com

Abstract: Right-censored length-biased data occurs frequently in observational studies and the probability of observing the data is proportional to the length of failure time. It causes great challenges to data analysis. Therefore, existing methods for traditional survival data cannon be used to length-biased data. Mean residual life models is an important characteristic of evaluating the remaining life of a subject having survived up to a given time. In this paper, we consider a flexible semiparametric varying-coefficient mean residual life model with length-biased data, in which some covariate coefficients are allowed to vary as functions of other variables. Making using of inverse probability method(IPW), we develop a three-step method to estimation, which can improve the efficiency of the estimation.

Also, both independent censoring and dependent censoring are considered. Furthermore, the asymptotic properties of the estimation are established. Simulation studies are conducted and shows that the proposed methods performs well. Finally, an illustrative example is given.

S084: Statistical Methods for Large-Scale Networks

Nonregular and Minimax Estimation of Individualized Thresholds in High dimension with Binary Responses

Yang Ning
Cornell University
E-mail: yn265@cornell.edu

Abstract: Given a large number of covariates Z, we consider the estimation of a high-dimensional parameter θ in an individualized linear threshold θ′Z for a continuous variable X, which minimizes the disagreement between sign(X−θ′Z) and a binary response Y. While the problem can be formulated into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to discontinuity of the sign function. Moreover, estimating θ even in the fixed-dimensional setting is known as a nonregular problem leading to nonstandard asymptotic theory. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter θ, we propose an empirical risk minimization approach based on a regularized smoothed loss function. The statistical and computational trade-off of the algorithm is investigated. Statistically, we show that the finite sample error bound for estimating θ in $\ell 2$ norm is $(s\log d/n)^{\beta/(2\beta+1)}$, where d is the dimension of θ, s is the sparsity level, n is the sample size and β is the smoothness of the conditional density of X given the response Y and the covariates Z. The convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. The Lepski's method is developed to achieve the adaption to the unknown sparsity s and smoothness β. Computationally, an efficient path-following algorithm is proposed to compute the solution path. We show that this algorithm achieves geometric rate of convergence for computing the whole path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis.

Machine Learning Methods For Estimation and Inference in Differential Networks

Mladen Kolar
University of Chicago
E-mail: mkolar@chicagobooth.edu

Abstract: We present a recent line of work on estimating differential networks and conducting statistical inference about parameters in a high-dimensional setting. First, we consider a Gaussian setting and show how to directly learn the difference between the graph structures. A debiasing procedure will be presented for construction of an asymptotically normal estimator of the difference. Next, building on the first part, we show how to learn the difference between two graphical models with latent variables. Linear convergence rate is established for an alternating gradient descent procedure with correct initialization. Simulation studies illustrate performance of the procedure. Finally, we will discuss how to do statistical inference on the differential networks when data are not Gaussian.

Network Clustering Hypothesis Testing

Junwei Lu
Harvard University
E-mail: junweilu@hsph.harvard.edu

Abstract: We propose a general framework to infer the clustering hypothesis based on stochastic block model. We build a likelihood ratio type of statistic and use symmetry property to clusters to infer the structure. We show our test is honest and powerful. We also establish the lower bound of general clustering testing problem.

Estimating Joint Latent Space Models for Network Data with High-Dimensional Node Variables

Xuefei Zhang
University of Michigan
E-mail: xfzhang@umich.edu

Abstract: Network latent space models assume each node is associated with an unobserved latent position in a Euclidean space, and such latent variables determine the probability of two nodes connecting with each other. In many applications, nodes in the network are often observed along with high-dimensional node variables. These node variables provide important information for understanding the network structure, however the classical network latent space models have several limitations for incorporating them. In this paper, we propose a joint latent space model where we assume that the latent variables not only explain the network structure, but also are informative for the multivariate node variables. We develop projected gradient descent algorithm that estimates the latent positions using a criterion incorporating both network structure and node variables. We establish theoretical properties of the estimators and provide insights on

how incorporating high-dimensional node variables information could improve the estimation accuracy of the latent positions. We demonstrate the improvement in latent variable estimation and the improvements in associated downstream tasks, such as node variables missing value imputation, by simulation and application to a Facebook data example.

S085: Leadership and Innovation in Drug Development Through Quantitative Research

Motivation: Statisticians in the pharmaceutical industry play an important role in promoting innovations in drug development across all phases, including discovery, early development, late development, and post-marketing. In this panel session, champions of innovation from different drug companies will share their visions and strategies for building strong infrastructure and organizations to advance and promote innovations in statistics and data science, as well as their views on key capabilities and competences needed to lead innovations for drug development.

S086: Real World Data and Evidence for Health Care Decision Making

Precision Health, Real World Data, and Artificial Intelligence Algorithms

Haoda Fu
Eli Lilly and Company
E-mail: fu_haoda@lilly.com

Abstract: Digital health is an important pharmaceutical industry trend in recent years, and it can bring significant disruptive innovation to transform healthcare industry. In this talk, we will provide an introduction on digital health and associated analytics challenges and opportunities. In particular, we will focus on the central role of personalized intervention in the era of digital health. ShortBio：Dr. Haoda Fu is a senior research advisor and a enterprise lead for Machine Learning, Artificial Intelligence, and Digital Connected Care from Eli Lilly and Company. Dr.Haoda Fu is a Fellow of ASA (American Statistical Association). He is also an adjunct professor of biostatistics department, Indiana university school of medicine. Dr. Fu received his Ph.D. in statistics from University of Wisconsin - Madison in 2007 and joined Lilly after that. Since he joined Lilly, he is very active in statistics methodology research. He has more than 90 publications in the areas, such as Bayesian adaptive design, survival analysis, recurrent event modeling, personalized medicine, indirect and mixed treatment comparison, joint modeling, Bayesian decision making, and rare events analysis. In recent years, his research area focuses on machine learning and artificial intelligence. His research has been published in various top journals including JASA, JRSS, Biometrics, ACM, IEEE, JAMA, Annals of Internal Medicine etc.. He has been teaching topics of machine learning and AI in large industry conferences including teaching this topic in FDA workshop. He was board of directors for statistics organizations and program chairs, committee chairs such as ICSA, ENAR, and ASA Biopharm session.

Real world data, machine learning and causal inference

Jie Chen
Merck Research Laboratory
E-mail: jie_chen@merck.com

Abstract: There has recently been an increasing interest in applying statistical and machine learning algorithms to real-world data (RWD) for causality assessment. One of the major challenges in analyzing RWD is confounding bias that can lead to spurious association between an exposure and an outcome. Although there are several different approaches (such as propensity score matching and stratification) to adjusting for confounding bias in causal inference, this presentation will focus on structural causal model (SCM) based machine learning methodologies including target learning and reinforcement learning and application of these approaches to RWD for causal inference.

Propensity Score Method to Adjust for Confounding in Observational Research: Progression, Challenges, and Opportunities

Zhong Yuan
Janssen Research & Development
E-mail: zyuan6@its.jnj.com

Abstract: Randomized clinical trial (RCT) remains as a gold standard research design to establish the benefits and risks of an intervention (e.g., drug or device), because randomization controls for measured and unmeasured confounding factors. However, the limitations associated with RCTs are also well recognized, including (but not limited to) economic issue, limited sample size, relatively short treatment duration, extensive inclusion and exclusion criteria, and underrepresentation of certain patient population. With increasingly available data sources in routine clinical practice, observational research offers a unique opportunity to evaluate effectiveness and safety of interventional therapies in the real-world setting, where the patient population may be much broader than what have been studied in clinical trials. Because bias and confounding are perceived as inherent limitations for this type of research, propensity score (PS) method is increasingly used to minimize such issues. In the current presentation, we will discuss that not all PS methodologies are created equal and highlight a couple of examples how PS method is applied in our observational research.

Multistate modeling and simulation of patient trajectories after allogeneic hematopoietic stem cell transplantation to inform drug development

Jiawei Wei
Novartis
E-mail: JIAWEI.WEI@NOVARTIS.COM

Abstract: We present a case study for developing clinical trial scenarios in a complex progressive disease with multiple events of interest. The idea is to first capture the course of the disease in a multistate Markov model, and then to simulate clinical trials from this model, including a variety of hypothesized drug effects. This case study focuses on the prevention of graft-versus-host disease (GvHD) after allogeneic hematopoietic stem cell transplantation (HSCT). The patient trajectory after HSCT is characterized by a complex interplay of various events of interest, and there is no established best method of measuring and/or analyzing treatment benefits. We characterized patient trajectories by means of multistate models that we fitted to a subset of the Center for International Blood and Marrow Transplant Research (CIBMTR) database. Events of interest included acute GvHD of grade III or IV, severe chronic GvHD, relapse of the underlying disease, and death. The transition probability matrix was estimated using the Aalen-Johansen estimator, and patient characteristics were identified that were associated with different transition rates. In a second step, clinical trial scenarios were simulated from the model assuming various drug effects on the background transition rates, and the operating characteristics of different endpoints and analysis strategies were compared in these scenarios.

This helped devise a drug development strategy in GvHD prevention after allogeneic HSCT. More generally, multistate models provide a rich framework for exploring complex progressive diseases, and the availability of a corresponding simulation machinery provides great flexibility for clinical trial planning.

S087: Innovative study designs and analyses for early-phase clinical trials

The paradox of increasing sample size and decreasing power in testing the difference between two independent binomial proportions

Youyi Fong
University of Washington, Department of Biostatistics
E-mail: youyifong@gmail.com

Abstract: Suppose we are to conduct a trial to compare the probabilities of a binary outcome between two populations. Let the number of subjects to be sampled from the two populations be denoted by m and n, respectively. Denote the number of successes to be observed in the two samples by a and c, respectively. Thus a is a binomial random variable binom(m,p1) and b is a binomial random variable binom(n,p2). We are interested in testing the null hypothesis p1=p2. This hypothesis testing problem has been studied extensively in the literature. It is generally agreed among practitioners that unconditional tests such as Boschloo's test (Boschloo 1970) and Barnard's CSM test (Barnard, 1947) are more powerful than conditional tests like Fisher's exact test. But what to make of this difference in power is not without controversy. For example, Barnard (1989) argued that "the various suggestions that have been made to increase the power of Fisher's test for 2 x 2 tables are shown to give no real increases." In this talk we seek to shed some light on this controversy by considering a case of increasing sample size and decreasing power.

Registration Enabling Seamless Phase 1/2 Oncology Trial Design

Jun Dong
Amgen Biopharmaceutical R&D (Shanghai) Co., Ltd.
E-mail: jund@amgen.com

Abstract: In recent years, cases have shown the possibility of regulatory (accelerated or conditional) approval with phase 2 data for oncology products, when the treatment effect is clinically meaningful and relatively large. This is also consistent with China CDE's policy for conditional approval where the drug meets urgent clinical needs. In this presentation, an example design is shown, where the phase 1 dose exploration, phase 1b dose expansion and phase 2 are seamlessly integrated. In the single arm phase 2 portion, the investigational product is compared with external control for efficacy evaluation. Conditional approval is then sought after with a phase 3 confirmatory trial. Other adaptive component can be incorporated as well to further accelerate the clinical development process.

Review and Examples of Master Protocols

Li Li
R&G Pharma
E-mail: li.li@rg-pharma.com

Abstract: In recent years, while traditional statistical methods create challenges in recruiting patients with rare genetic subtypes of a disease, master protocol is gaining more interest due to its capability to provide innovative/flexible solutions to expedite cancer drugs development. In this talk, we will review several different types of master protocols, including 'basket', 'umbrella' and 'platform' trials. We will introduce the features of these designs, discuss the advantages and disadvantages and comments on their implementation. We will also present several examples, and discuss the challenges encountered in application.

S088: Statistical Methods for Genomic and Transcriptomic Data Analysis

Detecting Dense and Sparse Signals in Omics Studies

Chi Song
The Ohio State University
E-mail: song.1188@osu.edu

Abstract: In omics studies, there is a common scenario when we are interested in testing whether there exists any signal of unknown proportion in a dataset that contains tremendous amount of noise. For example, in chromosome copy number detection problem, the goal is to detect the unknown proportion of carriers of certain copy number change among non-carriers; and in GWAS, we are interested in detecting an unknown proportion of SNPs in a SNP set that may associate with a certain disease. In this talk, I will propose a novel approach that adaptively combines statistical tests to detect both dense and sparse signals in high dimensional data. Simulation will be used to explore the properties of this approach around the theoretical detection boundary. I will also discuss the application of this new approach in SNP set analysis and microbial community analysis.

Dynamic Correlation Analysis for Omics Data

Tianwei Yu
Emory University
E-mail: tianwei.yu@emory.edu

Abstract: In high-throughput data, dynamic correlation between genes, i.e. changing correlation patterns under different biological conditions, can reveal important regulatory mechanisms. Given the complex nature of dynamic correlation, and the underlying conditions for dynamic correlation may not manifest into clinical observations, it is difficult to recover such signal from the data. Current methods seek underlying conditions for dynamic correlation by using certain observed genes as surrogates, which may not faithfully represent true latent conditions. In this study we develop a new method that directly identifies strong latent signals that regulate the dynamic correlation of many pairs of genes, named DCA: Dynamic Correlation Analysis. At the center of the method is a new metric for the identification of gene pairs that are highly likely to be dynamically correlated, without knowing the underlying conditions of the dynamic correlation. We validate the performance of the method with extensive simulations. In real data analysis, the method reveals novel latent factors with clear biological meaning, bringing new insights into the data.

Statistical model for background mutation rate in cancer genomes

Lin Hou
Tsinghua University
E-mail: houl@tsinghua.edu.cn

Abstract: The identification of driver genes is an important problem in cancer genome analysis. Driver genes are identified via hypothesis testing procedures, which contrasts observed counts with the background mutation rate. In this work, we introduce a statistical model to estimate the background mutation rate in cancer genomes. We are able to identify driver genes of relatively low prevalence with high precision. The proposed

method is evaluated in simulation settings and in real data.

scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition

Ruibin Xi
Peking University
E-mail: ruibinxi@math.pku.edu.cn

Abstract: Single cell RNA-sequencing (scRNA-seq) technology is developing at a fast pace and provides a higher resolution for single cell gene profiling, which enables us to better understand biological process at the single cell level. Nevertheless, scRNA-seq data suffers from a significant issue of down sampling and is thus biased by excessive zeros. While some of them are essential zeros indicating no expression, others are missing values known as dropouts due to insufficient amount of mRNA transcripts. To identify dropout zeros and replace them with underlying true expression levels, we develop scRMD, a statistical method for imputation in scRNA-seq data by borrowing information across genes and cells. It is shown that scRMD is able to produce both accurate and robust imputed results that lead to better downstream statistical analyses including detection of differential expression and clustering analysis.

S089: Dimension reduction with applications

Sufficient dimension reduction via random-partitions for the large-p-small-n problem

Su-Yun Huang
Academia Sinica
E-mail: syhuang@stat.sinica.edu.tw

Abstract: Sufficient dimension reduction (SDR) continues to be an active field of research. When estimating the central subspace (CS), inverse regression based SDR methods involve solving a generalized eigenvalue problem, which can be problematic under the large-p-small-n situation. In recent years, new techniques have emerged in numerical linear algebra, called randomized algorithms or random sketching, for high-dimensional and large scale problems. To overcome the large-p-small-n SDR problem, we combine the idea of statistical inference with random sketching to propose a new SDR method, called integrated random-partition SDR (iRP-SDR). Our method consists of the following three steps: (i) Randomly partition the covariates into subsets to construct an envelope subspace with low dimension. (ii) Obtain a sketch of the CS by applying a conventional SDR method within the constructed envelope subspace. (iii) Repeat the above two steps many times and integrate these multiple sketches to form the final estimate of the CS. After describing the details of these steps, the asymptotic properties of iRP-SDR are established. Unlike existing methods, iRP-SDR does not involve the determination of the structural dimension until the last stage, which makes it more adaptive to a high-dimensional setting. The advantageous performance of iRP-SDR is demonstrated via simulation studies and a practical example analyzing EEG data.

Robust linear discriminant analysis based on gamma-divergence

Ting-Li Chen
Academia Sinica
E-mail: tlchen@stat.sinica.edu.tw

Abstract: Linear discriminant analysis (LDA), a traditional method in linear classification, is widely used in pattern recognition and dimension reduction. In many practical cases, some data points are mislabeled, which may badly affect the LDA results. In this talk, we will first introduce gamma-divergence which is a more robust dispersion measure than the widely used Kullback-Leibler divergence. Based on minimum gamma-divergence, we proposed a more robust LDA type method. Instead of sample mean and sample covariance derived by minimum K-L divergence, weighted sample mean and weighted sample covariance from gamma-divergence can successfully reduce the effects of incorrect labels. We will analyze the robustness of our proposed method via its influence function. In the end, we will demonstrate the strength of our proposed method by simulation studies and applications on face data sets.

An adaptive clustering for curve data

Heng-Hui Lue
Tunghai University
E-mail: hhlue@thu.edu.tw

Abstract: We propose a new adaptive approach for clustering curve data. The data-adaptive searching method based on dimension reduction theory is proposed for estimating the basis functions and the sufficient dimension reduction space of predictors. These estimates are obtained through local linear approximation techniques without requiring a prespecified parametric model. A K-means clustering method is then adopted for curve clustering analysis. Several examples are reported for illustration.

Online Learning for Multiclass Classification with Applications

Henry Horng-Shing Lu
NCTU
E-mail: hslu@stat.nctu.edu.tw

Abstract: This talk will discuss the new developments of online learning for classification of multiple categories with various applications.

S090: Advances in survival analysis in the era of data science

Survival Analysis of Two-Level Hierarchical Clustered Data

Weijing Wang
National Chiao Tung U
E-mail: wjwang@stat.nctu.edu.tw

Abstract: A new model is proposed for right-censored survival data with multilevel clustering based on the hierarchical Kendall copula model with Archimedean clusters. This model easily accommodates clusters of unequal size and multiple clustering levels, without any structural conditions on the parameters or on the copulas used at various levels of the hierarchy. A step-wise estimation procedure is proposed and shown to yield consistent and asymptotically Gaussian estimates under mild regularity conditions. The model fitting is based on multiple imputation, given that the censoring rate increases with the level of the hierarchy. To check the model assumption of Archimedean dependence, a goodness-of test is developed. The finite-sample performance of the proposed estimators and of the goodness-of-fit test is investigated through simulations. The new model is applied to data from the study of chronic granulomatous disease.

Semiparametric regression analysis for length-biased and interval-censored data with a cure fraction

Chyong-Mei Chen
National Yang-Ming University
E-mail: cmchen2@ym.edu.tw

Abstract: Left-truncated (LT) data are often encountered in epidemiological cohort studies, where individuals are recruited according to a certain cross-sectional sampling criterion. Length-biased data, a special case of LT data, assume that the incidence of the initial event follows a

homogeneous Poisson process. In this article, we consider analysis of length-biased and interval-censored (LBIC) data with a nonsusceptible fraction. We first point out the importance of a well-defined target population, which depends on the priori knowledge for the support of the failure times of susceptible individuals. Given the appropriate target population, we can proceed a length-biased sampling and draw valid inferences from sample. When there is no covariate, we show that to maximize the full likelihood function, it suffices to consider discrete version of the survival function for the susceptible individuals with jump points at left-end points of the censoring intervals. Based on this result, we propose an EM algorithm to obtain the nonparametric maximum likelihood estimates (NPMLEs) of nonsusceptible rate and the survival function of the susceptible individuals. We also develop a novel graphical method for assessing the stationarity assumption. When covariate is present, we consider analyzing LBIC data under the Cox proportional hazards model with a nonsusceptible fraction, where the probability of being susceptible is determined by the logistic regression model. We construct the full likelihood function and obtain the NPMLEs of the regression parameters by employing the EM algorithm. The large sample properties of the NPMLEs are established. The performance of the NPMLE is assessed by simulations. The metabolic syndrome data are analyzed to illustrate our method.

Semiparametric copula-based analysis for treatment effects in the presence of treatment switching

Yi-Hau Chen
Academia Sinica
E-mail: yhchen@webmail.stat.sinica.edu.tw

Abstract: In controlled trials, “treatment switching” occurs when patients in one treatment group switch to the alternative treatment during the trial, and poses challenges to evaluation of the treatment effects owing to crossover of the treatments groups. In this work, we assume that treatment switches occur after some disease progression event, and view the progression and death events as two semicompeting risks. The proposed model consists of a copula model for the joint distribution of time-to-progression (TTP) and overall survival (OS) before the earlier of the two events, as well as a conditional hazard model for OS subsequent to progression. The copula model facilitates assessing the marginal distributions of TTP and OS separately from the association between the two events, and, in particular, the treatment effects on TTP and on OS in the absence of treatment switching. The proposed conditional hazard model for death subsequent to progression allows us to assess the treatment switching (crossover) effect on OS given occurrence of progression and covariates. General semiparametric transformation models are employed in the marginal models for TTP and OS. A nonparametric maximum likelihood procedure is developed for model inference, which is verified through asymptotic theory and simulation studies. The proposed analysis is applied to a lung cancer dataset to illustrate its real utility.

A nonparametric approach to semi-competing risks via causal mediation modeling

Yen-Tsung Huang
Academia Sinica
E-mail: yentsung@gmail.com

Abstract: The semi-competing risk problem arises when one is interested in the effect of an exposure or treatment on both intermediate (e.g., having cancer) and primary events (e.g., death) where the intermediate event may be censored by the primary event, but not vice versa. Here we propose a nonparametric approach casting the semi-competing risks problem in the framework of causal mediation modeling. We set up a mediation model with the intermediate and primary events, respectively as the mediator and the outcome, and define indirect effect (IE) as the effect of the exposure on the primary event mediated by the intermediate event and direct effect (DE) as that not mediated by the intermediate event. A Nelson-Aalen type of estimator with time-varying weights is proposed for direct and indirect effects where the counting process at time t of the primary event N_(2n_1) (t) and its compensator A_(n_1) (t) are both defined conditional on the status of the intermediated event right before t, N_1 (t^-)=n_1. We show that N_(2n_1) (t)-A_(n_1) (t)is a zero-mean martingale. Based on this, we further establish the asymptotic unbiasedness, consistency and asymptotic normality for the proposed estimators. Numerical studies including simulation and data application are presented to illustrate the finite sample performance and utility of the proposed method.

S092: Dynamic Design of Optimal Treatment Regimes

Optimal dynamic treatment regimes using decision lists

Yichi Zhang
University of Rhode Island
E-mail: yichizhang@uri.edu

Abstract: A dynamic treatment regime (DTR) formalizes precision medicine as a series of functions over decision points. At each decision point, it takes the available information of a patient as input and outputs a recommended treatment for that patient. A high quality DTR tailors treatment decisions to individual patient as illness evolves, and thus improves patient outcomes while reducing cost and treatment burden. To facilitate meaningful information exchange during the development of DTRs, it is important that the estimated DTR be interpretable in a subject matter context. We propose a simple, yet flexible class of DTRs whose members are representable as a short list of if-then statements. DTRs in this class are immediately interpretable and are therefore appealing choices for broad applications in practice. We develop a nonparametric Q-learning procedure to estimate the optimal DTR within this class. We establish its consistency and rate of convergence. We demonstrate the performance of the proposed method using simulations and a clinical dataset.

Improved doubly robust estimation in learning optimal individualized treatment rules

Yinghao Pan
University of North Carolina at Charlotte
E-mail: ypan8@uncc.edu

Abstract: Individualized treatment rules (ITRs) recommend treatment according to patient characteristics. There is a growing interest in developing novel and efficient statistical methods in constructing ITRs. We propose an improved doubly robust estimator of the optimal ITRs. The proposed estimator is based on a direct optimization of an augmented inverse-probability weighted estimator (AIPWE) of the expected clinical outcome over a class of ITRs. The method enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent when either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among the class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. Simulation studies show that the estimated ITRs obtained from our method yield better results than those

obtained from current popular methods. Data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study is analyzed as an illustrative example.

Subagging for Inference of the Mean Outcome Under Optimal Treatment Regimes

Chengchun Shi
NC State University
E-mail: cshi4@ncsu.edu

Abstract: Precision medicine is an emerging medical approach that allows physicians to select the treatment options based on individual patient information. The goal of precision medicine is to identify the optimal treatment regime (OTR) that yields the most favorable clinical outcome. Although considerable research has been devoted to estimating the optimal treatment regime (OTR) in the literature, less attention has been paid to statistical inference of the OTR. In this paper, we develop a novel inference method for the mean outcome under an OTR (the optimal value function) based on subsample aggregating (subagging) and refitted cross-validation. The proposed method can be applied to multi-stage studies where treatments are sequentially assigned over time.

Bootstrap aggregating (bagging) and subagging have been recognized as effective variance reduction techniques to improve unstable estimators or classifiers (Buhlmann and Yu, 2002). However, it re-mains unknown whether these approaches can yield valid inference results. We show the proposed confidence interval (CI) for the optimal value function achieves nominal coverage even in the nonregular cases where the OTR is not uniquely defined. In addition, due to the variance reduction effect of subagging, our method enjoys certain statistical optimality. Specifically, we prove the length of the proposed CI is on average shorter than the CI constructed based on the online one-step method (Luedtke and van der Laan, 2016). Moreover, under certain conditions on the propensity score function, we show the proposed CI is asymptotically narrower than the CI of the "oracle" method which works as well as if an OTR were known. Numerical studies are conducted to back up our theoretical findings.

Online experiment design for mapping large-scale neural circuits

Shizhe Chen
University of California, Davis
E-mail: szdchen@ucdavis.edu

Abstract: We consider high throughput circuit mapping experiments where subthreshold, postsynaptic responses of one neuron are recorded using whole-cell patch clamp, and optical stimulation is used to stimulate multiple genetically modified neurons per trial. In these experiments, we are interested in (i) inferring which neurons have synaptic connections with the patched neuron, and (ii) the properties of the presynaptic neurons. However, the amount of data one can collect is paltry compared to the extent of neural circuits because the preparations are short-lived. In addition, the patched neuron's responses are subject to intrinsic stochasticity due to the low spatial resolution of the optical stimulation and the biological variability in the responses of individual neurons to the optical stimulation. We propose an online procedure that automatically designs future trials during the experiment. Our procedure first focuses on detecting and eliminating disconnected cells with multi-spot stimulations, then learns properties of the connected cells with precise single-spot stimulations. To this end, we develop a robust method for fitting a physiobiologically plausible model for the observed postsynaptic events, which is used to learn the properties of the few connected cells. We derive a simplified working model that is fast to fit, which is used to detect the many disconnected cells.

S094: Statistical Methods in Complex Data Analysis

A maximum average power test for large scale time-course data of counts with applications to RNA-Seq analysis

Wen Zhou
Colorado State University
E-mail: wzhou70@asu.edu

Abstract: Experiments that longitudinally collect RNA sequencing (RNA-seq) data can provide transformative insights in biology research by revealing dynamic patterns of genes. Such experiments create great demands for new analytic approaches to identify differentially expressed (DE) genes based on large-scale time-course count data. Existing methods, however, are sub-optimal with respect to power and may lack theoretical justification. Furthermore, most existing tests are designed to distinguish among conditions based on overall differential patterns across time, though in practice, a variety of composite hypotheses are of more scientificc interest. Lastly, some current methods may fail to control the false discovery rate (FDR). In this paper, we propose a new model and testing procedure to address the above issues simultaneously. Specifically, conditional on a latent Gaussian mixture with evolving means, we model the data by negative binomial distributions. Motivated by Storey (2007) and Hwang and Liu (2010), we introduce a general testing framework based on the proposed model and show that the proposed test enjoys the optimality property of maximum average power. The test allows not only identification of traditional DE genes but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and demonstrate its good performance via simulation studies. The procedure reveals interesting biological insights when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of the marine diatom Phaeodactylum tricornutum.

Distributed Dual Averaging Variational Inference

Shiyuan He
Renmin University of China
E-mail: heshiyuan@ruc.edu.cn

Abstract: Many modern machine learning algorithms rely on Bayesian probabilistic models whose posterior is difficult to compute. Variational Inference~(VI) is a popular rescue for the intractable posterior. Despite its popularity, VI lacks the scalability to huge dataset, especially when the dataset is distributed stored on distinct servers. In this work, we presents a simple-to-implement yet efficient algorithm to tackle the problem. The algorithm depends on distributed dual averaging, and only the cumulant natural gradient is exchanged among servers to ensure global convergence. The algorithm is applicable to a wide range of Bayesian models, and only assumes the variational distribution belongs to the exponential family. Moreover, it exploits the Riemannian geometry of the exponential family for efficient update. We test the algorithm over several real world datasets, and demonstrate its increased speed of convergence.

A nonparametric Bayesian approach to simultaneous subject and cell heterogeneity discovery for single cell RNA-seq data

Xiangyu Luo
Renmin University of China

E-mail: xiangyuluo@ruc.edu.cn

Abstract: The advent of the single cell sequencing era opens new avenues for the personalized treatment. The first but important step is discovering the subject heterogeneity at the single cell resolution. We address the two-level-clustering problem of simultaneous subject subgroup discovery (subject level) and cell type detection (cell level) based on the single cell RNA sequencing (scRNA-seq) data from multiple subjects. However, current approaches either cluster cells without considering the subject heterogeneity or group subjects not using the single cell information. We develop a solid nonparametric Bayesian model SCSC (Subject and Cell clustering for Single-Cell data) to achieve subject and cell grouping at the same time without pre-specifying the subject subgroup number or the cell type number. An efficient blocked Gibbs sampler is then proposed for the posterior inference. The simulation study and the real application demonstrate the good performance of our model.

Broadcasted Nonparametric Tensor Regression

Kejun He
Renmin University of China
E-mail: kejunhe@ruc.edu.cn

Abstract: In this talk, we propose a broadcasted model to study the problem of nonlinear regressions with tensor covariates. The curse of dimensionality is tamed by simultaneously utilizing the low-rank tensor structure and broadcasting a uni-dimensional function within each component. With a regularized estimation, the proposed model shows the advantages of improved prediction performance and suggesting the important regions on the tensor covariates. A novel result on the upper and lower bounds for the eigenvalues of the spline basis matrix is derived in this paper to develop the asymptotic theory. We use both synthetic and real data sets to evaluate the empirical performance of the proposed broadcasted nonparametric regression model with some comparison methods, and the results confirm our theoretical findings.

S095: Nonparametric or semiparametric inference on complicated data

Bayesian Analysis of Semiparametric Hidden Markov Models with Latent Variables

Jingheng Cai
Sun Yat-sen University
E-mail: caijheng@mail.sysu.edu.cn

Abstract: In psychological, social, behavioral, and medical studies, hidden Markov models (HMMs) have been extensively applied to the simultaneous modeling of heterogeneous observation and hidden transition in the analysis of longitudinal data. However, the majority of the existing HMMs are developed in a parametric framework without latent variables. This study considers a novel semiparametric HMM, which comprises a semiparametric latent variable model to investigate the complex interrelationships among latent variables and a nonparametric transition model to examine the linear and nonlinear effects of potential predictors on hidden transition. The Bayesian P-splines approach and Markov chain Monte Carlo methods are developed to estimate the unknown functions and parameters. Penalized expected deviance, a Bayesian model comparison statistic, is employed to conduct model comparison. The empirical performance of the proposed methodology is evaluated through simulation studies. An application to a data set derived from the National Longitudinal Survey of Youth is presented.

Penalized generalized empirical likelihood with a diverging number of general estimating equations for censored data

Xiaodong Yan
Shandong University
E-mail: yanxiaodong@sdu.edu.cn

Abstract: This article considers simultaneous variable selection and parameter estimation as well as hypothesis testing in censored regression models with unspecified parametric likelihood. For the problem, we utilize certain growing dimensional general estimating equations and propose a penalized generalized empirical likelihood using the folded concave penalties. We first construct general estimating equations attaining the semiparametric efficiency bound with censored regression data and then establish the consistency and oracle properties of the penalized generalized empirical likelihood estimators. Furthermore, we show that the penalized generalized empirical likelihood ratio test statistic has an asymptotic standard central chi-squared distribution. The conditions of local and restricted global optimality of weighted penalized generalized empirical likelihood estimators are also discussed. We present an two-layer iterative algorithm for efficient implementation, and rigorously investigate its convergence property. The good performance of the proposed methods are demonstrated by extensive simulation studies and a real data example is provided for illustration.

A Model-averaging method for high-dimensional regression with missing responses at random

Niansheng Tang
Yunnan University
E-mail: nstang@ynu.edu.cn

Abstract: This article considers the ultrahigh-dimensional prediction problem in the presence of missing responses at random. A two-step model averaging procedure is proposed to improve prediction accuracy of conditional mean of response variable. The first step is to specify several candidate models, each with low-dimensional predictors. To implement this step, a new feature screening method is developed to distinguish from the active and inactive predictors via the inverse probability weighted rank correlation (IPWRC), and candidate models are formed by grouping covariates with similar size of IPWRC values. The second step is to develop a new criterion to find the optimal weights for averaging a set of candidate models via the weighted delete-one cross-validation (WDCV). Under some regularity conditions, we show that the proposed new screening statistic enjoys ranking consistency property, and the WDCV criterion asymptotically achieves the lowest possible prediction loss. Simulation studies and an example are illustrated by the proposed methodologies.

S096: Statistical inference on missing or censored data

Copula-based semiparametric analysis for time series data with detection limits

Yanlin Tang
East China Normal University
E-mail: yanlintang2018@163.com

Abstract: The analysis of time series data with detection limits is challenging due to the high-dimensional integral involved in the likelihood. Existing methods are either computationally demanding or rely on restrictive parametric distributional assumptions. We propose a semiparametric approach, where the temporal dependence is captured by parametric copula while the marginal distribution is estimated

nonparametrically. Utilizing the properties of copulas, we develop a new copula-based sequential sampling algorithm, which provides a convenient way to calculate the censored likelihood. Even without full parametric distributional assumptions, the proposed method still allows us to efficiently compute the conditional quantiles of the censored response at a future time point, and thus construct both point and interval predictions. We establish the asymptotic properties of the proposed pseudo maximum likelihood estimator, and demonstrate through simulation and the analysis of a water quality data that the proposed method is more flexible and leads to more accurate predictions than Gaussian-based methods for non-normal data.

Multiply Robust Subgroup Identification for Longitudinal Data with Dropouts

Guoyou Qin
Fudan University
E-mail: gyqin@fudan.edu.cn

Abstract: Subgroup identification serves as an important step towards precision medicine which has attracted great attention recently. On the other hand, longitudinal data with dropouts often arises in medical research. However there is little work in subgroup identification considering this data type. Therefore, in this paper we propose a new subgroup identification method based on concave fusion penalization and median regression for longitudinal data with dropouts. In order to deal with missingness, we introduce multiply robust weights which allow multiple models for the probability of being observed. As long as one of the models is correctly specified, the proposed estimator is able to achieve oracle property in the case of missingness. Furthermore, we develop an efficient algorithm and propose a modified Bayesian information criterion to select penalization parameter. The asymptotic properties of the proposed method is established under some regularity conditions. The numerical performance is illustrated in simulations and the proposed method is applied to the quality of life data from a breast cancer trail.

Bayesian Generalized Method of Moments Analysis for Complex Surveys

Puying Zhao
Department of statistics, Yunnan University
E-mail: pyzhao@ynu.edu.cn

Abstract: We consider Bayesian generalized method of moments inference with complex survey data where the finite population parameters are defined through the census estimating equations. The posterior distribution is formulated under the framework of generalized method of moments. We systematically evaluate large sample properties of the posterior density with any fixed or shrinking priors under the design-based framework. We show that the posterior distribution has the same shape of the quasi-log-likelihood function induced from the generalized method of moments quadratic function when the prior distribution is noninformative. Our results are valid under general unequal probability sampling designs with very mild conditions on the estimating functions and have major advantages on parameters defined through nonsmooth estimating functions. An effective Markov Chain Monte Carlo algorithm is developed to compute the proposed Bayes estimator and Bayesian credible intervals. Simulation results demonstrate that the proposed method works remarkably well for finite samples.

A Vine Copula Approach for Regression Analysis of Bivariate Current Status Data with Informative Censoring

Huiqiong Li
Yunnan University
E-mail: lihuiqiong@ynu.edu.cn

Abstract: Bivariate current status data occur in many areas and many authors have discussed their analysis and proposed many inference procedures (Hu et al., 2017; Jewell et al., 2005; Wang et al., 2015). However, most of these methods are for the situation where the observation or censoring is non-informative and sometimes one may face informative censoring (Chen et al., 2012; Ma et al., 2015; Zhang et al., 2005), where one has to deal with three correlated random variables. In this paper, a vine copula approach is developed for regression analysis of bivariate current status data in the presence of informative censoring. The proposed estimators are shown to be strongly consistent and the asymptotic normality and efficiency of the estimated regression parameter are also established. Numerical results suggest that the proposed method works well in practice.

S097: Recent Development on Missing Data Issues under Estimand Framework

Estimands, Missing Data, and Sensitivity Analysis

Geert Molenberghs
I-BioStat, Hasselt University & KU Leuven
E-mail: geert.molenberghs@uhasselt.be

Abstract: The presentation sets out by considering estimands, a very important topic in clinical trials. A connection is made with the much older use in survey sampling theory. Using an example from surrogate marker evaluation, it is discussed where information comes from: data, design, and assumptions. The latter may be unverifiable, hence the need to perform sensitivity analysis.

The setting is then broadened to various forms of enrichment; that is, every situation where the model contains more aspect than the data are able to provide information about. Subsets of the enrichment class are: (a) coarsening, where some data could have been observed but were not (e.g., missing data); (b) augmentation, where models, for convenience of interpretation, are supplemented with unobservables, such as random effects. The focus is then placed on incomplete data for the rest of the presentation.

A general framework for missing data is given, starting from Rubin's seminal work. The defining and transforming role of the National Academy of Sciences report from 2010 about the "Prevention and Treatment of Missing Data in Clinical Trials" is evocated. It is argued that the role of the patient should not be forgotten, next to academe, regulators, and industry.

It is shown that for every MAR model, there is a family of MNAR models that exhibits the same fit to the data. Hence, one cannot show that MAR holds or not, solely depending on the data. The implications for standard and sensitivity analyses are discussed.

Regarding standard analysis, the roles of ignorable likelihood and Bayesian analysis, multiple imputation, and inverse probability weighting (e.g., weighted generalized estimating equations), are discussed.

The presentation concludes with some brief illustrations of sensitivity analysis.

Cox Regression with Survival-Dependent Missing Covariate Values

Jun Shao
East China Normal University
E-mail: shao@stat.wisc.edu

Abstract: Analysis with time-to-event data in clinical and epidemiological

studies often encounters missing covariate values and the missing at random assumption is commonly adopted, which assumes that missingness depends on the observed data, including the observed outcome which is the minimum of survival and censoring time. However, it is conceivable that in certain settings, missingness of covariate values is related to the survival time but not to the censoring time. This is especially so when covariate missingness is related with an unmeasured variable affected by patient's illness and prognosis factors at baseline. If this is the case, then the covariate missingness is not at random as the survival time is censored, and it creates a challenge in data analysis. In this article, we propose an approach to deal with such survival-time-dependent covariate missingness based on the well known Cox proportional hazard model. Our method is based on inverse propensity weighting with the propensity estimated by nonparametric kernel regression. Our estimators are consistent and asymptotically normal, and their finite-sample performance is examined through simulation. An application to a real-data example is included for illustration.

The Challenges of Analyzing Drug Safety Data with Competing Risk Events and Some Thoughts

Aileen Zhu
China Novartis Institutes for BioMedical Research Co., Ltd.
E-mail: aileen.zhu@novartis.com

Abstract: Drug sponsors are often requested to do an investigation of serious safety events, such as cardiovascular events, in clinical trials. However, such an analysis is often hampered by the presence of competing risk events, e.g., non-event related death, that preclude the observation of the event types of interest. The competing risk events are especially of concern when the event rates are not balanced between active and control arms. We were recently requested by a health authority to address such a concern, with the suggested subdistribution proportional hazard model. However, this model only considers patients experiencing competing risk events to remain in the risk set, without investigating how likely these patients could have experienced the events of interest if they had not had the competing risk events. The yielded results being liberal for the arm with more competing risk events, often do not address health authority's concerns. In this presentation, two alternative approaches are proposed. First, the tipping point analysis, which can help find out until which point along the increase in the number of events among patients who experienced competing risk events in both arms, the conclusion of no treatment difference is altered. This approach can be extended to include also patients who dropped out. Second, a mixture model, by considering patients with an event of interest prior to study completion or not in two separate distributions. A gamma distribution is assumed for patients with an event of interest before study completion, and a logistic regression model is fitted to indicate whether patients had an event of interest or not. The estimation is later on used to impute the data for patients who had a competing risk event and who dropped out. The final estimation is then based on the multiply imputed data from the mixture model, which accounts for the uncertainty of whether an event of interest could have occurred for those who dropped out early (due to competing risk events or not).

Mixture of multivariate t linear Mixed Models With Missing Information

Tzy-Chy Lin
Center for Drug Evaluation
E-mail: tclin323@cde.org.tw

Abstract: Linear mixed-effects (LME) models have been widely used for longitudinal data analysis as it can account for both fixed and random effects, while simultaneously incorporating the variation on both within and between subjects. In clinical trials, some drugs may be more effective in Westerners than the Orientals. In this situation, such heterogeneity can be modeled by a finite mixture of LME models. The classical modeling approach for random effects and the errors parts are assumed to follow the normal distribution. However, normal distribution is sensitive to outliers and intolerance of outliers may greatly affect the model estimation and inference.

In this presentation, we propose a robust approach called the mixture of multivariate t LME models with missing information. To facilitate the computation and simplify the theoretical derivation, two auxiliary permutation matrices are incorporated into the model for the determination of observed and missing components of each observation. We describe a flexible hierarchical representation of the considered model and develop an efficient Expectation-Conditional Maximization Either (ECME) algorithm for carrying out maximum likelihood estimation. Simulation results and real data analysis are provided to illustrate the performance of the proposed methodology.

S098: Advances in sufficient dimension reduction and its applications

A Model - - free Conditional Screening Approach via Suf ficient Dimension Reduction

Meggie Wen
Missouri University of Science and Technology
E-mail: wenx@mst.edu

Abstract: Conditional variable screening arises when researchers have prior information regarding the importance of certain predictors, such as treatment effects in biological studies and market risk factors in fi nancial studies. It is natural to consider feature screening methods co nditioning on these known important predictors. Barut et al. (2016) pr oposed conditional sure independence screening (CSIS) to address this issue under the context of generalized linear models. While CSIS ou tperforms the marginal screening method when few of the factors are known to be important and/or significant correlations between some o f the factors exist, unfortunately, CSIS is model based and might fail when the models are misspecified. We propose a model-free conditio nal screening method under the framework of sufficient dimension reduction for ultrahigh-dimensional statistical problems. Numerical stud ies show our method easily beats CSIS for nonlinear models, and per forms comparable to CSIS for (generalized) linear models. Sure scree ning consistency property for our method is also proved.

Simultaneous estimation for semi-parametric multi-index models

Wenbo Wu
University of Texas at San Antonio
E-mail: wuwenbouoregon@gmail.com

Abstract: Estimation of a general multi-index model comprises determining the number of linear combinations of predictors (structural dimension) that are related to the response, estimating the loadings of each index vector, selecting the active predictors and estimating the underlying link function. These objectives are often achieved sequentially at different

stages of the estimation process. In this study, we propose a unified estimation approach under a semi-parametric model framework to attain these estimation goals simultaneously. The proposed estimation method is more efficient and stable than many existing methods where the estimation error in the structural dimension may propagate to the estimation of the index vectors and variable selection stages. A detailed algorithm is provided to implement the proposed method. Comprehensive simulations and a real data analysis illustrate the effectiveness of the proposed method.

Matching Using Sufficient Dimension Reduction for Causal Inference

Yeying Zhu
University of Waterloo
E-mail: yeying.zhu@uwaterloo.ca

Abstract: To estimate causal treatment effects, we propose a new matching approach based on the reduced covariates obtained from sufficient dimension reduction. Compared with the original covariates and the propensity score, which are commonly used for matching in the literature, the reduced covariates are nonparametrically estimable and are effective in imputing the missing potential outcomes, under a mild assumption on the low-dimensional structure of the data. Under the ignorability assumption, the consistency of the proposed approach requires a weaker common support condition. In addition, researchers are allowed to employ different reduced covariates to find matched subjects for different treatment groups. We develop relevant asymptotic results and conduct simulation studies as well as real data analysis to illustrate the usefulness of the proposed approach.

S099: Recent Advance in Bayesian Data Science

Bayesian Variable Selection with Application to High Dimensional EEG Data

Dipak Dey
Faculty
E-mail: dipak.dey@uconn.edu

Abstract: Due to the immense technological advances, very often we encounter data in high-dimensions. Any set of measurements taken at multiple time points for multiple subjects leads to data of more than two dimensions (matrix of covariates for each subject). In this talk, we present a Bayesian method for binary classification of subject-level responses by building binary regression models using latent variables along with the well-known spike and slab priors. We also study the scaled normal priors on the parameters, as they cover a large family of distributions. Due to the computational complexity, we build many local (at different time points) models and make predictions using the temporal structure between the local models. We perform variable selection for each of these local models. If the variables are locations, then the variable selection can be interpreted as spatial clustering. We show the results of a simulation study and also present the performance of these models on multi-subject neuroimaging (EEG) data.

Registration Enabling Seamless Phase 1/2 Oncology Trial Design

Lei Cao
Changchun University of Technology
E-mail: caol661@nenu.edu.cn

Abstract: Motivated from the AppRise data, we develop the autoregressive models for the item response and the response time jointly. One of the novelty of the proposed models is to account for the dependence for each of the item response and the response time as well as the dependence between the item response and the response time due to the sequential nature of the AppRise data. In addition, a new deviance information criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML) are constructed by integrating out subject-specific ability parameters and speed parameters. We further derive novel decompositions of DIC and LPML into two components, namely, the marginal DIC and LPML and the conditional DIC and the conditional LPML. These new conditional DIC or the conditional LPML can be used to assess the gain in the fit of the response data by using the response time data. We further compute the concordance measures for the response data as well as the response time data. These concordance measures are further used to confirm the gain in the fit of the response data jointly modeled with the response time. A detailed analysis of the AppRise data is carried out to demonstrate the applicability and usefulness of the proposed methodology.

Bayesian Hierarchical Spatial Regression Models for Spatial Data in the Presence of Missing Covariates with Applications

Zhihua Ma
Shenzhen University
E-mail: mazh1993@outlook.com

Abstract: In many applications, survey data are collected from different survey centers in different regions. It happens that in some circumstances, response variables are completely observed while the covariates have missing values. In this paper, we propose a joint spatial regression model for the response variable and missing covariates via a sequence of one-dimensional conditional spatial regression models. We further construct a joint spatial model for missing covariate data mechanisms. The properties of the proposed models are examined and a Markov chain Monte Carlo sampling algorithm is used to sample from the posterior distribution. In addition, the Bayesian model comparison criteria, the modified Deviance Information Criterion (mDIC) and the modified Logarithm of the Pseudo-Marginal Likelihood (mLPML), are developed to assess the fit of spatial regression models for spatial data. Extensive simulation studies are carried out to examine empirical performance of the proposed methods. We further apply the proposed methodology to analyze a real data set from a Chinese Health and Nutrition Survey (CHNS) conducted in 2011.

Bayesian Meta-Regression Hierarchical Models for Cholesterol Data

Ming-Hui Chen
University of Connecticut
E-mail: ming-hui.chen@uconn.edu

Abstract: A flexible class of multivariate meta-regression models are proposed for Individual Patient Data (IPD). The methodology is well motivated from 26 pivotal Merck clinical trials that compare statins (cholesterol lowering drugs) in combination with ezetimibe and statins alone on treatment-naïve patients and those continuing on statins at baseline. The research goal is to jointly analyze the multivariate outcomes, Low Density Lipoprotein Cholesterol (LDL-C), High Density Lipoprotein Cholesterol (HDL-C), and Triglycerides (TG). These three continuous outcome measures are correlated and shed much light on a subject's lipid status. The proposed multivariate meta-regression models allow for different skewness parameters and different degrees of freedom for the multivariate outcomes from different trials under the general class of

skewed t-distributions. The theoretical properties of the proposed models are examined and an efficient Markov chain Monte Carlo (MCMC) computational algorithm is developed for sampling from the posterior distribution under the proposed multivariate meta-regression model. In addition, the Conditional Predictive Ordinates (CPOs) are computed via an efficient Monte Carlo method. Consequently, the logarithm of the pseudo marginal likelihood and Bayesian residuals are obtained for model comparison and assessment, respectively. A detailed analysis of the IPD meta data from the 26 Merck clinical trials is carried out to demonstrate the usefulness of the proposed methodology. This is a joint work with Sung Duk Kim, Joseph G. Ibrahim, Arvind Shah, and Jianxin Lin.

S100: New Advances on Statistical Modeling of Complex Data

Mixtures of factor analysis models with covariates for multiply censored dependent data

TSUNG-I LIN
National Chung Hsing University
E-mail: tilin@nchu.edu.tw

Abstract: Censored data arise frequently in diverse applications in which observations to be measured may be subject to some upper and lower detection limits due to the restriction of experimental apparatus such that they are not exactly quantifiable. Mixtures of factor analyzers with censored data (MFAC) have been recently proposed for model-based density estimation and clustering of high-dimensional data under the presence of censored observations. We consider an extended version of MFAC with covariates to accommodate multiply censored dependent variables and develop two analytically feasible EM-type algorithm for computing maximum likelihood estimates of the parameters with closed-form expressions. Moreover, we provide an information-based method to compute asymptotic standard errors of mixing proportions and regression coefficients. The utility and performance of our proposed methodologies are illustrated through two real data examples.

Analysis of Multivariate Longitudinal Data with Censored and Intermittent Missing Responses

WAN-LUN WANG
E-mail: wlunwang@fcu.edu.tw

Abstract: The multivariate linear mixed model (MLMM) has emerged as an important analytical tool for longitudinal data with multiple outcomes. However, the analysis of multivariate longitudinal data could be complicated by the presence of censored measurements because of a detection limit of the assay in combination with unavoidable missing values arising when subjects miss some of their scheduled visits intermittently. This paper presents a generalization of the MLMM approach, called the MLMM-CM, for a joint analysis of the multivariate longitudinal data with censored and intermittent missing responses. A computationally feasible expectation maximization-based procedure is developed to carry out maximum likelihood estimation within the MLMM-CM framework. Moreover, the asymptotic standard errors of fixed effects are explicitly obtained via the information-based method. The proposed methodology is demonstrated through a simulation and a case study from an AIDS clinical trial. Experimental results reveal that our method is able to provide more satisfactory performance as compared with the traditional MLMM approach.

Bayesian Analysis of Survival Data with Missing Censoring Indicators

Mauricio Castro
Pontificia Universidad Catolica de Chile
E-mail: mcastro@mat.uc.cl

Abstract: In some large clinical studies, it may be impractical to perform the physical examination to every subject at his/her last monitoring time in order to diagnose the occurrence of the event of interest. This gives rise to survival data with missing censoring indicators where the probability of missing may depend on time of last monitoring and some covariates. We present a fully Bayesian semi-parametric method for such survival data to estimate regression parameters of Cox's proportional hazards model (Cox, 1972). Theoretical investigation and simulation studies show that our method performs better than competing methods. We apply the proposed method to data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study.

Likelihood-based Inference for Mixed-Effects Models with Censored Response Using Skew-Normal Distribution

Victor Hugo Lachos Davila
University of Connecticut
E-mail: hlachos@uconn.edu

Abstract: Mixed-effects models are commonly used to fit longitudinal or repeated measures data. A complication arises when the response is censored, for example, due to limits of quantification of the assay used. Although normal distributions are commonly assumed for random effects, such assumption may be unrealistic obscuring important features of among-individual variation. We relax this assumption by consider a likelihood-based inference for linear and nonlinear mixed effects models with censored response (NLMEC/LMEC) based on the multivariate skew-normal distribution. An ECM algorithm is developed for computing the maximum likelihood estimates for NLMEC/LMEC with the standard errors of the fixed effects and the exact likelihood value as a by-product. The algorithm uses closed-form expressions at the E-step, that rely on formulas for the mean and variance of a truncated multivariate skew-normal distribution. The proposed algorithm is implemented in the R package skewlmec(). It is applied to analyze longitudinal HIV viral load data in two recent AIDS studies. In addition, a simulation study is conducted to examine the performance of the proposed methods.

S101: New Advance in Bayesian Approach for Complex Data

A score-based two-stage Bayesian network method for detecting causal SNPs

Yue Zhang
Shanghai Jiao Tong University
E-mail: yue.zhang@sjtu.edu.cn

Abstract: With the development of genome-wide association studies, how to gain more meaningful information from the volumes of data has become an issue of common concern, especially when dealing with the problems like identifying the epistatic interactions associated with complex diseases. The huge amount of possible combinations of all SNPs makes the task difficult. Thus, developing powerful and robust methods for detecting epistatic interactions is of great importance. In this paper, we propose a score-based two-stage Bayesian network method to identify genomewide epistatic interactions in a case-control design. This method combines the ideas of constraint-based methods and score-and-search methods to learn

the structure of the disease-centred local Bayesian network. We compare our new algorithm with several common epistasis interactions detecting methods in simulation studies. The results show that our method has a good accuracy and stability. Besides, its successful application on SMRI dataset suggests that our algorithm has the ability to handle real GWAS data.

High-dimensional posterior consistency for hierarchical non-local priors in regression

Xuan Cao
University of Cincinnati
E-mail: xuan.cao@uc.edu

Abstract: The choice of tuning parameters in Bayesian variable selection is a critical problem in modern statistics. In particular, for Bayesian linear regression with non-local priors, the scale parameter in the non-local prior density is an important tuning parameter which reflects the dispersion of the non-local prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero. Current approaches treat the scale parameter as given, and suggest choices based on prior coverage/asymptotic considerations. In this talk, we consider the fully Bayesian approach with the pMOM non-local prior and an appropriate Inverse-Gamma prior on the tuning parameter to analyze the underlying theoretical property. Under standard regularity assumptions, we establish strong model selection consistency in a high- dimensional setting, where p is allowed to increase at a polynomial rate with n or even at a sub-exponential rate with n. Through simulation studies, we demonstrate that our model selection procedure can outperform other Bayesian methods which treat the scale parameter as given, and commonly used penalized likelihood methods, in a range of simulation settings.

Bayesian Spatially Dynamic Variable Selection for Spatial Point Process

Jieying Jiao
University of Connecticut
E-mail: jieying.jiao@uconn.edu

Abstract: Poisson point process is widely used to study the relationship between occurrence of events in space and spatial covariates. Variable selection problem of spatial point process model with spatially varying coeffcients have not yet received much attention. The spike-slab prior has been universally used for Bayesian variable selection. To capture spatially varying variable selection uncertainty, we introduce a new spatially dynamic spike-slab prior for spatial point process model. Several theoretical results are examined in this paper. An efficient Markov chain Monte Carlo algorithm is developed for our proposed methods. Extensive simulation studies are carried out to show the effeteness of our proposed methods. The usefulness of our model is illustrated by an application in BCI data.

Bayesian Spatial Heterogeneity Pursuit Regression Models

Guanyu Hu
University of Connecticut
E-mail: guanyu.hu@uconn.edu

Abstract: Most existing spatial clustering literatures discussed the cluster algorithm for spatial responses. In this paper, we consider a nonparametric Bayesian clustered regression in order to detect clusters in the covariate effects. Our proposed method is based on the geographically weighted Chinese restaurant process which provides a probabilistic framework for simultaneous inference of the number of clusters and the clustering configurations. A Markov chain Monte Carlo sampling algorithm is used to sample from the posterior distribution of the proposed model. In addition, Bayesian model diagnostic techniques are developed to assess the fitness of our proposed model, and check the accuracy of clustering results. Extensive simulation studies are conducted to evaluate the empirical performance of the proposed models. For illustration, our methodology is applied to a housing cost dataset of Georgia.

S102: Structure and correlation analysis

A Consistent Independence Test via Projected Mutual Info rmation

Yaowu Zhang
Shanghai University of Finance and Economics
E-mail: zhangyaowucp@163.com

Abstract: We propose a nonparametric independence test based on m utual information. Distinguished from the previous work, we estimate the mutual information in a conditional density form, whose dimensio n could be reduced to 1 with novel projection methods. The optimal projection direction, which we name as maximum unit direction, is es timated by maximizing a penalized mutual information. An independe nce test is later on carried out via the newly estimated mutual inform ation and is shown to be insensitive to the dimensions. The test is c onsistent against all global alternatives, and can detect local alternativ es at a fast rate as if the model is univariate. Numerical results indic ate that the test is more powerful compared with other existing indep endence tests, especially when the sample size is small or the dimens ion is large.

Test of Independence via Categorically Weighted Distance Correlation

Wei Zhong
Xiamen University
E-mail: wzhong@xmu.edu.cn

Abstract: It is of particular importance to understand the relationship among random variables in statistical inference. In this paper, motivated by the classification problems, we place focus on a test of independence between a categorical random variable and a random vector. A new Categorically Weighted Distance Correlation (CWDC) is developed to measure the dependence between a categorical random variable and a random vector. Asymptotical distributions and associated theoretical properties of the new CWDC-based test of independence are studied. The test is robust to distribution assumptions and outliers. Monte Carlo simulations demonstrate its excellent finite-sample performance.

Best Subset Selection in Linear, Logistic and CoxPH Models

Canhong Wen
University of Science and Technology of China
E-mail: wench@ustc.edu.cn

Abstract: We introduce a new R package, BeSS, for solving the best subset selection problem in linear, logistic and Cox's proportional hazard (CoxPH) models. It utilizes a highly efficient active set algorithm based on primal and dual variables, and supports sequential and golden search strategies for best subset selection. We provide a C++ implementation of the algorithm using Rcpp interface. We demonstrate through numerical experiments based on enormous simulation and real datasets that the new BeSS package has competitive performance compared to other R packages for best subset selection purpose.

S103: Selective Inference and Multiple Comparisons

Selective Inference after Unsupervised Hidden-Structure Identification

Ichiro Takeuchi
Nagoya Institute of Technology
E-mail: takeuchi.ichiro@nitech.ac.jp

Abstract: In the past few years, a new statistical inference framework for data-driven hypotheses called post-selection inference (PSI; also a.k.a. selective inference) has been actively studied. PSI framework enables the evaluation of data-driven hypotheses by taking into account that hypotheses were selected by applying a complex algorithm to complex data. Although the target of PSI framework has been mostly limited to feature selection problems so far, we have recently extended the framework to unsupervised learning scenarios and demonstrated that statistical reliability of unsupervised learning results can be properly evaluated. In this talk, we introduce PSI methods for evaluating the statistical reliability of data-driven hypotheses obtained by clustering and segmentation algorithms and illustrate the advantages by applying the methods to biomedical data.

Perturbation of the expected Minkowski functional and its applications

Satoshi Kuriki
The Institute of Statistical Mathematics
E-mail: kuriki@ism.ac.jp

Abstract: The Minkowski functional is a series of geometric quantities including the volume, the surface area, and the Euler characteristic. In this talk, we consider the Minkowski functional of the excursion set (sup-level set) of an isotropic smooth random field on arbitrary dimensional Euclidean space. Under the setting that the random field has weak non-Gaussianity, we provide the perturbation formula of the expected Minkowski functional. This result is a generalization of Matsubara (2003) who treated the 2- and 3-dimensional cases. The Minkowski functional is used in astronomy and cosmology as a test statistic for testing Gaussianity of the cosmic microwave background (CMB), and to characterize the large-scale structures of the universe. Besides, the expected Minkowski functional of the highest degree is the expected Euler-characteristic of the excursion set, which approximates the upper tail probability of the maximum of the random field. This methodology is referred to as the Euler-characteristic method (the expected Euler-characteristic heuristic), and is used in multiple testing problems. We explain some applications of the perturbation formulas in these contexts. (Joint work with Takahiko Matsubara)

Adjusting the bias of bootstrap probability with "negative" sample size and its applications to clustering and multiple comparisons

Hidetoshi Shimodaira
Kyoto University / RIKEN AIP
E-mail: shimo@i.kyoto-u.ac.jp

Abstract: Bootstrap resampling has been widely used for calculating confidence level of discrete decisions. The bootstrap probability of a specified hypothesis is calculated as the frequency of observing the same hypothesis in bootstrapped datasets. Although the bootstrap probability can be interpreted as a Bayesian posterior probability, it is biased as a frequentist p-value caused by the "curvature" of the boundary surface of the hypothesis in the parameter space (Efron and Tibshirani 1998). This bias is successfully adjusted via the multiscale bootstrap (Shimodaira 2004) by utilizing the scaling-law of the bootstrap probability in "m-out-of-n" bootstrap. It is very surprising that the unbiased p-value is obtained by extrapolating the normalized bootstrap probability to m = -n. By now, the method has been widely used in clustering (pvclust) and phylogenetics (CONSEL). We review the method, and also its extension to selective inference (Terada and Shimodaira 2017).

Selective inference for the problem of regions via multiscale bootstrap with applications to clustering and regression

Yoshikazu Terada
Osaka University
E-mail: terada@sigmath.es.osaka-u.ac.jp

Abstract: A general approach to selective inference is considered for hypothesis testing of the null hypothesis represented as an arbitrary shaped region in the parameter space of multivariate normal model. This approach is useful for hierarchical clustering where confidence levels of clusters are calculated only for those appeared in the dendrogram, thus subject to heavy selection bias. Our computation is based on a raw confidence measure, called bootstrap probability, which is easily obtained by counting how many times the same cluster appears in bootstrap replicates of the dendrogram. We adjust the bias of the bootstrap probability by utilizing the scaling-law in terms of geometric quantities of the region in the abstract parameter space, namely, signed distance and mean curvature. Although this idea has been used for non-selective inference of hierarchical clustering, its selective inference version has not been discussed in the literature. Our bias-corrected p-values are asymptotically second-order accurate in the large sample theory of smooth boundary surfaces of regions, and they are also justified for nonsmooth surfaces such as polyhedral cones. The p-values are asymptotically equivalent to those of the iterated bootstrap but with less computation. The proposed algorithm is applied to practical selective inference problems on hierarchical clustering and regression.

S104: Model Selection and Information Criteria

A Cp Criterion for Semiparametric Causal Inference

Yoshiyuki Ninomiya
The Institute of Statistical Mathematics
E-mail: ninomiya@ism.ac.jp

Abstract: For marginal structural models, which play an important role in causal inference, we consider a model selection problem within a semiparametric framework using inverse-probability-weighted estimation or doubly robust estimation. In this framework, the modelling target is a potential outcome that may be missing, so there is no classical information criterion. We define a mean squared error for treating the potential outcome and derive an asymptotic unbiased estimator as a Cp criterion using an ignorable treatment assignment condition. Simulation shows that the proposed criterion outperforms a conventional one by providing smaller squared errors and higher frequencies of selecting the true model in all the settings considered. Moreover, in a real-data analysis we found a clear difference between the two criteria.

High-dimensionality-adjusted Consistent Information Criterion in Multivariate Linear Models

Hirokazu Yanagihara
Hiroshima University
E-mail: yanagi-hiro@hiroshima-u.ac.jp

Abstract: In this paper, we deal with a variable selection in multivariate linear regression models, based on minimization of the generalized Cp

(GCp) criterion when the dimension of the response variables vector may be large. Recently, Yanagihara (2016) proposed the high - dimensionality - adjusted consistent GCp (HCGCp) that is the consistent GCp criterion for which consistency can be achieved whenever the dimension of the response variables vector is fixed or goes to infinity. A high probability of selecting the true subset of explanatory variables can be expected under a moderate sample size when the HCGCp criterion is used to select variables, even when there is a high - dimensional response variables vector. Unfortunately, Yanagihara (2016) showed the consistency of HCGCp under the assumption that the true distribution of response variables is the multivariate normal distribution. Needless to say, nobody knows the true distribution of response variables. Hence, we show that the robustness to nonnormality of the HCGCp, i.e., the HCGCp has a consistency property under the violation of normality of the true distribution.

Convergence rate of importance weighted orthogonal greedy algorithm

Shinpei Imoti
Hiroshima University
E-mail: imori@hiroshima-u.ac.jp

Abstract: This paper studies a variable selection problem under the covariate shift when the number of covariates is larger than the sample size. The orthogonal greedy algorithm (OGA) is an effective variable selection procedure for such a high-dimensional situation. However, its validity may lose under the covariate shift. Thus, by considering the covariate shift, we propose an importance weighted OGA as an extension of OGA, and derive its convergence rate with respect to prediction error.

Risk-estimation based predictive densities for heteroskedastic hierarchical models

Keisuke Yano
The University of Tokyo
E-mail: yano@mist.i.u-tokyo.ac.jp

Abstract: We consider the problem of estimating the predictive density in a heteroskedastic Gaussian model under general divergence loss. Based on a conjugate hierarchical set-up, we consider generic classes of shrinkage predictive densities with both location and scale hyper-parameters. For any $alpha$-divergence loss, we propose a risk-estimation based methodology for tuning these shrinkage hyper-parameters. Our proposed predictive density estimators enjoy optimal asymptotic risk properties that are in concordance with the optimal shrinkage calibration point estimation results established by Xie, Kou, Brown (2012) for heteroskedastic hierarchical models. These $alpha$-divergence risk optimality properties of our proposed predictors are not shared by empirical Bayes predictive density estimators that are calibrated by traditional methods such as by maximizing the likelihood or by using method of moments. We conduct several numerical studies to compare the non-asymptotic performance of our proposed predictive density estimators with other competing methods and obtain encouraging results.

S105: Statistical Theory for Neural Networks and Machine Learning

Generalization error of deep learning and its learning dynamics from compression ability point of view

Taiji Suzuki
The University of Tokyo
E-mail: taiji@mist.i.u-tokyo.ac.jp

Abstract: One of biggest issues in deep learning theory is its generalization ability despite the huge model size. The classical learning theory suggests that overparameterized models cause overfitting. However, practically used large deep models avoid overfitting, which is not well explained by the classical approaches. To resolve this issue, several attempts have been made. Among them, the compression based bound is one of the promising approaches. In this talk, we give a new frame-work for compression based bounds. The bound gives even better rate than the one for the compressed network by improving the bias term. We can obtain a data dependent generalization error bound which gives a tighter evaluation than the data independent ones. Moreover, we discuss the learning dynamics of deep learning and how a compressible network is trained using the notion of neural tangent kernel.

Fisher information of deep neural networks with random weights

Ryo Karakida
National Institute of Advanced Industrial Science and Technology (AIST)
E-mail: karakida.ryo@aist.go.jp

Abstract: Investigating deep neural networks (DNNs) with random weights has given promising results in both theory and practice. When such random DNNs are sufficiently wide, we can formulate their behavior by using simple analytical equations through coarse-graining of the random weights. In this talk, we briefly overview recent advances on the random DNNs and adopt them to the analysis of the Fisher information matrix (FIM). We reveal that the usual setting of wide DNNs leads to pathological distortion of the FIM's eigenvalue spectrum. In particular, we show that the FIM has pathologically large eigenvalues and they determine a learning rate necessary for gradient methods to converge. Our FIM's statistics also provide suggestions to deep learning methods such as batch normalization and neural tangent kernel.

Generalization Analysis for Mechanism of Deep Learning via Nonparametric Statistics

Masaaki Imaizumi
The Institute of Statistical Mathematics
E-mail: imaizumi@ism.ac.jp

Abstract: We theoretically investigate an advantage of deep neural networks (DNNs) which empirically perform better than other standard methods. While DNNs have empirically shown higher performance than other methods, understanding its mechanism is still a challenging problem. From an aspect of the nonparametric statistics, it is known many standard methods attain the optimal rate of errors for standard settings such as smooth functions, and thus it has not been straightforward to find theoretical advantages of DNNs. Our study fills this gap by extending a class for data generating processes. We mainly consider the following two points; non-smoothness of functions and intrinsic structures of data distributions. We derive the generalization error of estimators by DNNs with a ReLU activation, and show that convergence rates of the generalization error can describe an advantage of DNNs over some of the other methods. In addition, our theoretical result provides guidelines for selecting an appropriate number of layers and edges of DNNs. We provide numerical experiments to support the theoretical results.

Statistical Inference with Unnormalized Models

Takafumi Kanamori

Tokyo Institute of Technology
E-mail: kanamori@c.titech.ac.jp

Abstract: Parameter estimation of unnormalized models is a challenging problem because normalizing constants are not calculated explicitly and maximum likelihood estimation is computationally infeasible. Although some consistent estimators have been proposed earlier, the problem of statistical efficiency does remain. In this talk, we propose a unified, statistically efficient estimation framework for unnormalized models and several novel efficient estimators with reasonable computational time.

S106: Dependent Data Analysis

Bayesian spatio-temporal modeling of Arctic sea ice extent

Bohai Zhang
Nankai University
E-mail: bohaizhang@nankai.edu.cn

Abstract: Arctic sea ice extent has drawn considerable interest from geoscientists for the last two decades owing to its rapid decline. In this paper, we propose a Bayesian spatio-temporal hierarchical model for Arctic sea ice extent data, where a latent spatio-temporal Gaussian process is used to model the data dependence and linked to the observations, which here are binary. Through a simulation study, we investigate how parameter uncertainty in a complex hierarchical model can influence spatio-temporal prediction. These results inform how inference will proceed on Arctic sea ice extent over a period of more than twenty years. Covariates that are physically motivated are chosen through autologistic diagnostics. Finally, new summary statistics are proposed to detect the changing patterns of Arctic sea ice between successive time periods.

Autologistic network model on binary data for disease progressionstudy

Huiyan Sang
Texas A&M University
E-mail: huiyan@stat.tamu.edu

Abstract: This paper focuses on analysis ofspatiotemporal binary data with absorbing states. The research was motivated by a clinical study on amyotrophic lateral sclerosis (ALS), a neurological disease marked by gradual loss of muscle strength over time in multiple body regions. We propose an autologisticregression model to capture complex spatial and temporal dependencies in muscle strength among different muscles. As it is not clear how the disease spreads from one muscle to another, it may not be reasonable to define a neighborhood structure based on spatialproximity. Relaxing the requirement for prespecification of spatial neighborhoods as in existing models, our method identifies an underlying network structure empirically to describe the pattern of spreading disease. The model also allows the network autoregressiveeffects to vary depending on the muscles' previous status. Based on the joint distribution derived from this autologistic model, the joint transition probabilities of responses among locations can be estimated and the disease status can be predicted in thenext time interval. Model parameters are estimated through maximization of penalized pseudo - likelihood. Postmodel selection inference was conducted via a bias - correction method, for which the asymptotic distributions were derived. Simulation studies were conductedto evaluate the performance of the proposed method. The method was applied to the analysis of muscle strength loss from the ALS clinical study.

Integrative interaction analysis of multi-omics data

Mengyun Wu
Shanghai University of Finance and Economics
E-mail: wu.mengyun@mail.shufe.edu.cn

Abstract: For the etiology, progression, and treatment of complex diseases, gene-environment (G-E) interactions have important implications beyond the main G and E effects. With the small sample sizes of omics profiling studies and noisy nature of omics data, the interaction analysis on a single dataset often leads to unsatisfactory results. In recent profiling studies, a prominent trend is to collect measurements on multi-omics data, including gene expressions as well as their regulators (copy number alteration, microRNA, methylation, etc.) on the same subjects. In our study, we propose a joint interaction analysis approach in an integrative perspective based on the biclustering and regularized estimation techniques, uniquely effectively accommodating the regulation relationships among multi-omics data. Simulations show that the proposed approach has significantly improved identification performance. In the analysis of cancer multi-omics data, biologically sensible findings different from the alternatives are made.

S107: Recent Advances in Probability Theory and Related Fields

On Cramer-von Mises statistic for the spectral distribution of random matrices

Zhigang Bao
Hong Kong University of Science and Technology
E-mail: mazgbao@ust.hk

Abstract: Let F_n and F be the empirical and limiting spectral distributions of an n by n Wigner matrix. The Cramer-von Mises (CvM) statistic is a classical goodness-of-fit statistic that characterizes the distance between F_n and F in l^2-norm. In this talk, we will consider a mesoscopic approximation of the CvM statistic for Wigner matrices, and derive its limiting distribution. The distribution fits well the heuristic prediction given by the log-correlated Gaussian field approximation for the stochastic filed of F_n(t). This is a joint work with Yukun He.

Concentration Inequalities for Point Processes

Hanchao Wang
Shandong University
E-mail: wanghanchao@sdu.edu.cn

Abstract: There have been a lot of research activities around phenomena of measure concentration in the past decades. In this talk, I will introduce some results on concentration inequalities for point processes.

Crossing probabilities in 2D critical lattice models

Hao Wu
Tsinghua University
E-mail: hao.wu.proba@gmail.com

Abstract: The planar Ising model is one of the most studied lattice models in statistical physics. It was introduced in the 1920s by W. Lenz as a model for magnetic materials. R. Peierls showed in 1936, in two (and higher) dimensions, an order-disorder phase transition in fact occurs at a certain critical temperature. Ever since, there has been active research to understand the 2D Ising model at criticality, where it enjoys conformal invariance in the scaling limit. In this talk, we give crossing probabilities of multiple interfaces in the critical planar Ising model with alternating boundary conditions. Besides, we also explain that a similar formula on the crossing probabilities also holds for critical Percolation and level lines of Gaussian Free Field.

Gaussian unitary ensembles with pole singularities near the soft

edge and a system of coupled Painlevé XXXIV equations

Lun Zhang
School of Mathematical Sciences, Fudan University
E-mail: lunzhang@fudan.edu.cn

Abstract: In this talk, we consider the singularly perturbed Gaussian unitary ensembles defined by the measure begin{equation*} frac{1}{C_n} e^{- ntextrm{tr}, V(M;lambda,vec{t};)}dM, end{equation*} over the space of $n times n$ Hermitian matrices M, where $V(x;lambda,vec{t};):= 2x^2 + sum_{k=1}^{2m}t_k(x-lambda)^{-k}$ with $vec{t}= (t_1, t_2, ldots, t_{2m})in mathbb{R}^{2m-1} times (0,infty)$, in the multiple scaling limit where $lambdato 1$ %approaches the soft edge of the limiting spectrum of Gaussian unitary ensemble together with $vec{t} to vec{0}$ as $nto infty$ at appropriate related rates. We obtain the asymptotics of the partition function, which is described explicitly in terms of an integral involving a smooth solution to a new coupled Painlev'e system generalizing the Painlev'e XXXIV equation. The large n limit of the correlation kernel is also derived, which leads to a new universal class built out of the Psi-function associated with the coupled Painlev'e system. Joint work with Dan Dai ans Shuai-Xia Xu.

S108:Topics in survival and longitudinal analysis with applications to clinical studies

Empirical likelihood for additive hazards regression model with case II interval censored failure time data

Chunjie Wang
Changchun University of Technology
E-mail: cjwang2014@126.com

Abstract: Interval censored failure time data occur in many areas. Many approaches have been proposed under various hazards regression models based on the asymptotic normality in survival statistics studies. We proposed an empirical likelihood approach for an additive hazards model with case II interval censored failure time data. For a vector of regression parameters, an empirical log-likelihood ratio is defined and it is shown its limiting distribution is a standard chi-squared distribution. Finite sample performance of our proposed empirical likelihood approach are demonstrated by simulation studies, and it shows that the empirical likelihood method provides more accurate inference than the normal approximation method. Empirical likelihood appraoch is applied to analyzing a real study of the breast cancer data.

Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression

Hui Zhao
Zhongnan University of Economics and Law
E-mail: hzhao@mail.ccnu.edu.cn

Abstract: The simultaneous estimation and variable selection for Cox model has been discussed by several authors (Fan and Li, 2002; Huang and Ma, 2010; Tibshirani, 1997) when one observes right-censored failure time data. However, there does not seem to exist an established procedure for interval-censored data, a more general and complex type of failure time data, except two parametric procedures in Scolas et al. (2016) and Wu and Cook (2015). To address this, we propose a broken adaptive ridge (BAR) regression procedure that combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage. In particular, the method allows for the number of covariates to be diverging with the sample size. Under some weak regularity conditions, unlike most of the existing variable selection methods, we establish both the oracle property and the grouping effect of the proposed BAR procedure. We conduct an extensive simulation study and show that the proposed approach works well in practical situations and deals with the collinearity problem better than the other oracle-like methods. An application is also provided.

Functional Mixed Effects model for joint analysis of longitudinal and cross-sectional growth data

Yingchun Zhou
East China Normal University
E-mail: yczhou@stat.ecnu.edu.cn

Abstract: A new method is proposed to perform joint analysis of longitudinal and cross-sectional growth data to improve the efficiency of the estimates. Clustering is first performed to group similar subjects in cross-sectional data to form a pseudo longitudinal data set, then the pseudo longitudinal data and real longitudinal data are combined and analyzed by using a functional mixed effects model. To account for the variational difference between pseudo and real longitudinal growth data, it is assumed that the covariance functions of the random effects and the variance functions of the measurement errors for pseudo and real longitudinal data can be different. Various simulation studies and real data analysis demonstrate the good performance of the method.

S109: Statistical and Machine Learning Methods with Application in AI Transportation

Statistics, Optimization and Deep Learning in the ride-sharing Industry

Fan Zhou
School of Statistics and Management, Shanghai University of Finance and Economics
E-mail: zhoufan@mail.shufe.edu.cn

Abstract: In this talk, we introduce some fundamental questions people from ride-sharing industries are interested in. Firstly, we introduce a novel class of equilibrium metrics (EMs) to quantify spatio-temporal equilibrium of dynamic supply-demand networks defined on the same graph. The two key components of EMs are to formulate the spatio-temporal equilibrium problem as an unbalanced optimal transport problem and to develop an efficient linear programming algorithm to solve such transport problem. On the other hand, prediction of customer demands from each original location to a destination helps ride-sharing platforms to better understand their market mechanism. However, most existing prediction methods ignore the network structure of OD flow data and fail to utilize the topological dependencies among related OD pairs. In this paper, we propose a spatial-temporal origin-destination (STOD) model, with a novel convolutional neural network (CNN) filter to learn the spatial features of OD pairs from a graph perspective and an attention structure to capture their long-term periodicity.

A statistical and machine learning framework for new energy vehicle ride sharing system

Kaixian Yu
Didi Chuxing
E-mail: kaixiany@gmail.com

Abstract: Recently, the number of electric vehicles (EVs) served on the online ride-hailing companies, like Uber, Didi Chuxing, increased rapidly. Not like conventional fuel vehicles, EVs have some unique characteristics:

they do not travel as far as fuel vehicles, and it takes much longer for EVs to be charged. Adapting these characteristics into the dispatching system of online ride-hailing companies becomes increasingly important. In this talk, we will present our recent progress on two major components of an EV friendly dispatching system. Firstly, we will introduce a stochastic partial differential equation approach to model the power consumption by an EV. The power consumption model takes real time vehicle and environment factors into account to estimate the state of charge. Secondly, we will introduce a deep multi-objective reinforcement learning approach to solve the order dispatching problem based on the estimated state of charge of EVs. Some results on real data and simulated system will be shown as well.

Recent advances in landmark-based scalable spectral clustering

Guangliang Chen
San Jose State University
E-mail: guangliang.chen@sjsu.edu

Abstract: Spectral clustering has emerged as a very effective clustering approach; however, it is computationally very expensive. As a result, there has been considerable effort in the machine learning community to develop fast, approximate spectral clustering algorithms that are scalable to large data. Notably, most of those methods use a small set of landmark points selected from the given data. In this talk we present two new landmark-based scalable spectral clustering algorithms that are developed based on novel document-term and bipartite graph models. We demonstrate the superior performance of our proposed algorithms by comparing them with the state-of-the-art methods on some benchmark data sets. Finally, we provide a unified view of all the old and new landmark-based spectral clustering methods.

S111: Strategic and Statistical Considerations in Early Phase Drug Development

Proof of Concept Decision Making in Phase 1b Cohort Expansion Study

Chao Zhu
Eli Lilly and Company
E-mail: zhu_chao_sh@lilly.com

Abstract: Single-arm cohort expansion studies are increasingly used as proof-of-concept (POC) studies for immunotherapies as compared to conventional randomized POC studies. We proposed a Bayesian three-tier framework to facilitate quantitative decision making in early phase clinical development programs which potentially could better balance between development speed and uncertainties. Characteristics of the framework were evaluated via simulations and we compared our framework with other available frameworks.

Proof of Mechanism and Proof of Concept Clinical Trials

Naitee Ting

Boehringer Ingelheim.com

E-mail: naitee.ting@boehringer-ingelheim.com

Abstract: In typical clinical development programs for new treatment of chronic diseases, Phase I trials recruit healthy volunteers. Hence clinical efficacy cannot be observed in Phase I. The first Phase II clinical trial is known as the Proof of Concept (PoC) trial and this is to compare the test drug against a placebo control in patients with the target disease. However, in Phase I, biomarkers based on mechanism of action can be observed. Therefore some of the Phase I clinical trials are considered as the Proof of Mechanism (PoM) trials. There are huge difficulties and challenges in design and analysis of PoM and PoC clinical trials. This presentation proposes a communication tool to help statisticians in communicating various risks with non-statisticians regarding risks

Bayesian Basket trial Design Accounting for Multiple Cutoffs of the Ambiguous Biomarker

Jin Xu
East China Normal University
E-mail: jxu@stat.ecnu.edu.cn

Abstract: Basket trial design enrolls patients with different cancer types but the same genetic mutation or biomarker to evaluate the treatment effect of a targeted therapy. However, the explicit biomarker sometimes may not be clearly identified. In this article, we propose a Bayesian basket trial design to account for multiple cutoffs of ambiguous biomarkers and select the optimal cutoff to maximize the benefit subpopulation. A two-stage design is proposed for the estimation. Secondly, we propose a simple method to cluster homogeneous subgroups within the families defined by the biomarker to enhance the power for detecting efficacious subgroups. Extensive simulations are conducted to demonstrate the operating characteristics of the two estimation methods in terms of probability of correct selection of optimal cutoff and probability of efficacy.

Bayesian Model-Assisted Designs for the Easy Conduct and Efficient Design of Phase I/II Trials: Keep It Simple and Smart!

J. Jack Lee
University of Texas MD Anderson Cancer Center
E-mail: jjlee@mdanderson.org

Abstract: Many novel adaptive designs have been published in the last two decades. Although these novel adaptive designs possess good statistical properties, most of them have not been widely implemented in real trials. Three major impediments are (1) complicated statistical modelling, (2) demanding computations, and (3) hard-to-be-used by clinical researchers. We introduce a new class of novel adaptive designs, known as model-assisted designs, to remove these hurdles and to facilitate the increasing use of novel designs to improve the efficiency and success of Phase I/II trials. Model-assisted designs combine the transparency and simplicity of the conventional algorithm-based designs with the superiority and rigorousness of model-based designs. They enjoy the superior performance comparable to more complicated, model-based designs, but can be implemented as simple as the conventional designs. A few model-assisted designs will be discussed including the Bayesian optimal interval (BOIN) design, the time-to-event BOIN (TITE-BOIN), the BOIN combination design, and the Bayesian Optimal Phase 2 (BOP2) design for simple complex endpoints, etc. Similar to all trial designs, the design parameters need to be carefully chosen and calibrated through simulation studies to ensure desirable operating characteristics. Freely available Shiny applications are provided at www.trialdesign.org to facilitate the adoption of model-assisted designs. Model-assisted designs establish a new KISS principle: Keep It Simple and Smart!

S112: CWS Special Invited Session: Recent Advances in Statistical Methods for Genomic Data

Statistical Inference of Chromatin 3D Structures from DNA Methylation Data

Shili Lin

Ohio State University
E-mail: shili@stat.osu.edu

Abstract: It has been hypothesized that, in complex genomes, a gene may be controlled not only by regulatory elements binding to its promotor, but also by distal enhancers and repressors. Molecular techniques have been developed to detect physical contacts between distant genomic loci, which support the hypothesis and validate the theory that communications between such elements are achieved through spatial organization of chromosomes to bring genes and their regulatory elements into close proximity. Although typical data used to understand the 3D structure are Hi-C based, recent work has shown that DNA methylation data obtained from primary patient samples can effectively recover the A/B compartment, the hallmark feature of chromatin 3D structure. In this talk, I will describe a statistical inference procedure for understanding the chromatin 3D structure. I will then discuss its application to a low-grade glioma (LGG) dataset to dissect long-range chromatin interactions and structural differences between two group of LGGs. Our results support the value of DNA methylation for understanding 3D structures, which show clear compartmentalization of active and inactive chromatins.

Network hub-node prioritization of gene regulation with intra-network association

Chuhsing Kate Hsiao
National Taiwan University
E-mail: ckhsiao@ntu.edu.tw

Abstract: To identify and prioritize the influential hub genes in a gene-set or biological pathway, most analyses rely on calculation of marginal effects or tests of statistical significance. Such procedures may be inappropriate if dependence between gene nodes exists, and if the hub nodes require more attention than others. The dependence may manifest itself in correlation between the nodes or in the topology of the pathway network. The highly connected hub genes may play a more important role for the whole network to function properly. Here we develop a pathway activity score incorporating the local effect of gene nodes as well as intra-network affinity measures. This score summarizes the expression levels in a gene-set/pathway for each sample, with weights on local and network information, respectively. The score is next used to examine the impact of each node through a leave-one-out evaluation. Two cancer studies, one involving RNA-Seq from breast cancer patients with high-grade ductal carcinoma in situ and one microarray expression data from ovarian cancer patients, and simulation analysis are used to assess the performance of the procedure, and to compare with existing methods with/without consideration of correlation and network information. The identified hub genes have reproduced previous findings; some are currently undergoing clinical trials for target therapy. The results show that the proposed procedure can provide a useful and complementary list of recommendation for prioritizing causal hubs.

Gene-set Integrative Omics Analysis Using Tensor-based Association Tests

Jung-Ying Tzeng
NC State University
E-mail: jytzeng@ncsu.edu

Abstract: Integrative multi-omics analyses integrate complementary level of information from different molecular events and have great potentials to detect novel disease genes and elucidate disease mechanisms. One major focus of integrative analysis has been on identifying gene-sets associated with clinical outcomes, and a common strategy is to regress clinical outcomes on all genomic variables in a gene set. However, such joint modeling methods encounter the challenges of high-dimensional inference, especially the sample size is usually moderate either due to research resources or missing data. In this work, we consider a tensor-based framework to enhance model efficiency for variable-wise inference. The tensor framework reduces the number of parameters by accounting for the inherent matrix structure of an individual's multi-omics data and naturally incorporates the relationship among omics variables. We study the variable-specific testing procedure under tensor regression framework and enhance computational efficiency of the omics tensor modeling. We evaluate the performance of the tensor-based test using simulations and real data application on the Uterine Corpus Endometrial Carcinoma dataset from the Cancer Genome Atlas (TCGA).

Genetic factors selection for association study with imbalanced case-control samples

Charlotte Wang
Department of Mathematics, Tamkang University
E-mail: cwang@mail.tku.edu.tw

Abstract: Traditional statistical models for analyzing balanced case-control samples are successful by optimizing overall accuracy; however, class imbalance problems in case-control study become more and more common in real applications. Optimizing overall accuracy will result in high accuracy of the control group but low accuracy of the case group. In addition, the statistical models with large non-associated genetic factors (noise variables) will be biased and the control group (major class) will dominate the results of genetic factors selection. Hence, how to incorporate characteristic of the case group (minor group) to select associated genetic factors in imbalanced case-control samples becomes an important issue. In this talk, I will introduce a method for genetic factors selection in imbalanced case-control samples.

S113: Current Challenges in Functional Data Analysis

Spatially Dependent Functional Data: Covariance Estimation, Principal Component Analysis, and Kriging

Yehua Li
University of California at Riverside
E-mail: yehuali@ucr.edu

Abstract: We consider spatially dependent functional data collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. The functional response is the sum of a spatially dependent functional effect and a spatially independent functional nugget effect. Observations on each function are made on discrete time points and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. Under a unified framework for both sparse and dense functional data, where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. Advantages of the proposed approach over existing methods are demonstrated through simulation studies and a real data application to the

home price-rent ratio data in the San Francisco Bay Area.

Weak Separability Test for Spatial Functional Fields

Fang Yao
Peking University
E-mail: fyao@math.pku.edu.cn

Abstract: For spatially dependent functional data, a generalized Karh unen-Loeve expansion is commonly used to decompose data into an a dditive form of temporal components and spatially correlated coefficie nts. This structure provides a convenient tool to investigate the space-time interactions, but may not always hold for more complex spatial-f unctional data. In this paper, we introduce a concept of weak separab ility, and propose formal testing procedures to examine the validity of the general Karhunen-Loeve decomposition. Asymptotic distribution o f the test statistic is studied to avoid using resampling procedures, e.g. bootstrap. Both parametric and nonparametric approaches are develop ed to estimate the asymptotic covariance, by constructing lagged cova riance estimators. We demonstrate the efficacy of our methods via si mulations under settings of grid and non-grid data, and illustrate their applications using two examples: Harvard forest data and China PM2. 5 data.

On multiple segmentation of a functional data sequence

Jeng-Min CHIOU
Academia Sinica
E-mail: jmchiou@stat.sinica.edu.tw

Abstract: We propose a statistical approach to detecting changes in a sequence of functional data. We derive the global optimality criterion for the changepoints as a foundation to determine the segments using different segmentation schemes. The method is robust to the number of changepoints. The asymptotic distribution of the differentials between the objective functions can be used to judge the significance of a functional change. We demonstrate the proposed method through an application to multiple traffic segmentation and examine its performance through a simulation study. (This is a joint work with Yu-Ting Chen).

Wasserstein Gradients for the Temporal Evolution of Probability Distributions

Yaqing Chen
University of California, Davis
E-mail: yaqchen@ucdavis.edu

Abstract: Many studies have been conducted on flows of probability measures, often in terms of gradient flows. We utilize a generalized notion of derivatives with respect to time to model the instantaneous evolution of empirically observed one-dimensional distributions that vary over time and develop consistent estimates for these derivatives. Employing local Fréchet regression and working in local tangent spaces with regard to the Wasserstein metric, we derive the rate of convergence of the proposed estimators. The resulting time dynamics are illustrated with time-varying distribution data that include yearly income distributions and the evolution of mortality over calendar years.

S114: Highlights of Statistica Sinica

A Model-averaging method for high-dimensional regression with missing responses at random

Nianshen Tang
Yunnan University
E-mail: nstang@ynu.edu.cn

Abstract: This article considers the ultrahigh-dimensional prediction problem in the presence of missing responses at random. A two-step model averaging procedure is proposed to improve prediction accuracy of conditional mean of response variable. The first step is to specify several candidate models, each with low-dimensional predictors. To implement this step, a new feature screening method is developed to distinguish from the active and inactive predictors via the inverse probability weighted rank correlation (IPWRC), and candidate models are formed by grouping covariates with similar size of IPWRC values. The second step is to develop a new criterion to find the optimal weights for averaging a set of candidate models via the weighted delete-one cross-validation (WDCV). Under some regularity conditions, we show that the proposed new screening statistic enjoys ranking consistency property, and the WDCV criterion asymptotically achieves the lowest possible prediction loss. Simulation studies and an example are illustrated by the proposed methodologies.

Estimation of Sparse Functional Additive Models with Adaptive Group LASSO

Jiguo Cao
Simon Fraser University
E-mail: jiguo_cao@sfu.ca

Abstract: We study a flexible model to address the lack of fit in conventional functional linear regression models. This model, called the sparse functional additive model, is used to characterize the relationship between a functional predictor and a scalar response of interest. The effect of the functional predictor is represented in a nonparametric additive form, where the arguments are the scaled functional principal component scores. Component selection and smoothing are considered when fitting the model in order to reduce the variability and enhance the prediction accuracy, while providing an adequate fit. To achieve these goals, we propose using the adaptive group LASSO method to select relevant components and smoothing splines and, thus, obtain a smoother estimate of those relevant components. Simulation studies show that the proposed estimation method compares favorably with conventional methods in terms of prediction accuracy and component selection. Furthermore, the advantages of our estimation method are demonstrated using two real-data examples.

Understanding and Utilizing the Linearity Condition in Dimension Reduction

Masayuki Henmi
The Institute of Statistical Mathematics
E-mail: henmi@ism.ac.jp

Abstract: When using inverse regression methods in dimension reduction models, the popular linearity condition has a paradoxical e ect: ignoring the linearity condition yields a more e cient estimator than making use of the linearity condition. By considering classes of parametric models, which include the linearity condition as a special case, we examine this phenomenon using a geometrical approach, and provide an intuitive and extended explanation. Our findings explain what the real cause of the paradox is, indicate how to properly handle the linearity condition and reveal the true role of the linearity condition. Our analysis directly leads to new estimators that further improve the existing ecient estimator that did not specifically account for the linearity condition and the possible constant variance condition.

S115: Recent developments in discriminant and multivariate analysis

A Doubly-Enhanced EM Algorithm for Model-Based Tensor Clustering

Qing Mai
Florida State University
E-mail: mai@stat.fsu.edu

Abstract: Modern scientific studies often generate tensor data, which calls for innovative statistical analysis methods. An important problem is to perform tensor clustering to understand the heterogeneity in the data. Many existing clustering methods are based on the K-means clustering and ignore the correlation among features. We propose a model-based approach to enable probabilistic interpretation. Our statistical model leverages the tensor structure to reduce the number of parameters for parsimonious modeling. Moreover, our model explicitly exploits the correlation for better variable selection and clustering. We propose a doubly-enhanced EM (DEEM) algorithm to perform clustering under this model. Both the E-step and the M-step are carefully tailored for tensor data. Theoretical studies confirm that DEEM achieves consistent clustering even when the dimension of each mode of the tensors grow at an exponential rate of the sample size, while numerical studies demonstrate favorable performance of DEEM in comparison to existing methods.

Robust Principal Component Analysis by Manifold Optimization

Teng Zhang
University of Central Florida
E-mail: Teng.Zhang@ucf.edu

Abstract: Robust PCA is a widely used statistical procedure to recover a underlying low-rank matrix with grossly corrupted observations. This work considers the problem of robust PCA as a nonconvex optimization problem on the manifold of low-rank matrices, and proposes two algorithms (for two versions of retractions) based on manifold optimization. It is shown that, with a proper designed initialization, the proposed algorithms are guaranteed to converge to the underlying low-rank matrix linearly. Compared with a previous work based on the Burer-Monterio decomposition of low-rank matrices, the proposed algorithms reduce the dependence on the conditional number of the underlying low-rank matrix theoretically. Simulations and real data examples confirm the competitive performance of our method.

Linear discriminant analysis with high dimensional mixed variables

Binyan Jiang
The Hong Kong Polytechnic University
E-mail: by.jiang@polyu.edu.hk

Abstract: With the rapid development of modern measurement technologies, datasets containing both discrete and continuous variables are more and more commonly seen in different areas, and in particular, the dimensions of the discrete and continuous variables can oftentimes be very high. Though discriminant analysis for mixed variables under the traditional fixed dimension setting has been well studied since the 80's, promising approaches taking into account both the high dimensionality and the mixing nature of the data sets are still missing. In this paper, we aim to developing a simple yet useful classification rule that addresses both the high dimensionality and the mixing nature of the variables simultaneously. Our framework is built on a location model, under which we further propose a semiparametric formulation for the optimal Bayes rule. We show that the optimal classification direction and the intercept in the optimal rule can be estimated separately. Efficient direct estimation schemes are then developed to obtain consistent estimators of the discriminant components. Asymptotic results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by simulation and real data studies .

S116: Finding structures in complex data

On Consistency and Sparsity for Large-Scale Curve Time Series with Application to Autoregressions

Xinghao Qiao
London School of Economics
E-mail: x.qiao@lse.ac.uk

Abstract: Modelling a large collection of curve time series arises in a broad spectral of real applications. Under such a scenario, not only the number of functional variables, p, is large relative to the number of temporally dependent curve observations, n, but each curve itself is an infinite-dimensional object, posing a challenging task. In this talk, a standard three-step procedure is proposed to address such large-scale problems. To provide theoretical guarantees for the three-step procedure, we focus on multivariate stationary processes and propose a novel functional stability measure based on their spectral properties. Such stability measure facilitates the development of some useful concentration bounds on sample covariance matrix functions, which serve as a fundamental tool for further consistency analysis, in particular, for deriving rates of convergence on the regularized estimates in large p, small n settings. As functional principal component analysis (FPCA) is one of the key dimension reduction techniques in the first step, we also investigate the consistency properties of the relevant estimated terms under a FPCA framework. To illustrate with an important application, we consider vector functional autoregressive models and develop a regularization approach to estimate autoregressive coefficient functions under the sparsity constraint. Using our derived convergence results, we investigate the theoretical properties of the regularized estimate under high-dimensional scaling. Finally, the finite-sample performance of the proposed method is examined through both simulations and a public financial dataset.

High-dimensional principal component analysis with heterogeneous missingness

Tengyao Wang
London's Global University
H-mail: tengyao.wang@ucl.ac.uk

I-**Abstract:** We study the problem of high-dimensional Principal Component Analysis (PCA) with missing observations. In simple, homogeneous missingness settings with a noise level of constant order, we show that an existing inverse-probability weighted (IPW) estimator of the leading principal components can (nearly) attain the minimax optimal rate of convergence. However, deeper investigation reveals both that, particularly in more realistic settings where the missingness mechanism is heterogeneous, the empirical performance of the IPW estimator can be unsatisfactory, and moreover that, in the noiseless case, it fails to provide exact recovery of the principal components. Our main contribution, then, is to introduce a new method for high-dimensional PCA, called primePCA, that is designed to cope with situations where observations may be missing

in a heterogeneous manner. Starting from the IPW estimator, primePCA iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. It turns out that the interaction between the heterogeneity of missingness and the low-dimensional structure is crucial in determining the feasibility of the problem. We therefore introduce an incoherence condition on the principal components and prove that in the noiseless case, the error of primePCA converges to zero at a geometric rate when the signal strength is not too small. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism. Our numerical studies on both simulated and real data reveal that primePCA exhibits very encouraging performance across a wide range of scenarios.

Optimal nonparametric change point detection and localization

Yi Yu
University of Bristol
E-mail: y.yu@bristol.ac.uk

Abstract: We study change point detection and localization for univariate data in fully nonparametric settings in which, at each time point, we acquire an i.i.d. sample from an unknown distribution. We quantify the magnitude of the distributional changes at the change points using the Kolmogorov–Smirnov distance. We allow all the relevant parameters – the minimal spacing between two consecutive change points, the minimal magnitude of the changes in the Kolmogorov–Smirnov distance, and the number of sample points collected at each time point – to change with the length of time series. We generalize the renowned binary segmentation (e.g. Scott and Knott, 1974) algorithm and its variant, the wild binary segmentation of Fryzlewicz (2014), both originally designed for univariate mean change point detection problems, to our nonparametric settings and exhibit rates of consistency for both of them. In particular, we prove that the procedure based on wild binary segmentation is nearly minimax rate-optimal. We further demonstrate a phase transition in the space of model parameters that separates parameter combinations for which consistent localization is possible from the ones for which this task is statistical unfeasible. Finally, we provide extensive numerical experiments to support our theory. R code is available at https: //github.com/hernanmp/NWBS.

S117: New methods and theory for analysing Big Data

Least Squares Approximation for a Distributed System

Hansheng Wang
Peking University
E-mail: hansheng@pku.edu.cn

Abstract: In this work we develop a distributed least squares approximation (DLSA) method, which is able to solve a large family of regression problems (e.g., linear regression, logistic regression, Cox's model) on a distributed system. By approximating the local objective function using a local quadratic form, we are able to obtain a combined estimator by taking a weighted average of local estimators. The resulting estimator is proved to be statistically as efficient as the global estimator. In the meanwhile it requires only one round of communication. We further conduct the shrinkage estimation based on the DLSA estimation by using an adaptive Lasso approach. The solution can be easily obtained by using the LARS algorithm on the master node. It is theoretically shown that the resulting estimator enjoys the oracle property and is selection consistent by using a newly designed distributed Bayesian Information Criterion (DBIC). The finite sample performance as well as the computational efficiency are further illustrated by extensive numerical study and an airline dataset. The airline dataset is 52GB in memory size. The entire methodology has been implemented by Python for a de-facto standard Spark system.

Sparsifying Deep Neural Networks with Generalized Regularized Dual Averaging

Guang Cheng
Purdue Statistics
E-mail: chengg@purdue.edu

Abstract: Deep learning has shown strikingly good performance in image classification, machine translation, text-to-speech translation. However, as modern deep neural networks (DNNs) require huge computational resources to store and process, deploying DNNs on devices and systems, e.g. mobile devices, requires to address storage and computational constraints. It is commonly believed that DNN is usually overparametrized, and it is possible to shrink DNN without sacrificing its accuracy via, e.g. pruning (Han et al, 2015). However, pruning is not efficient as it requires pre- and post-training of the model, and there is no theoretical justification for it. In this talk, we introduce a generalization of regularized dual averaging (gRDA) for sparsifying DNN, which does not require pre- and post-training. Under infinitesimal learning rate, gRDA has the same learning trajectory as stochastic gradient descent (SGD). Therefore, asymptotically gRDA achieve the same generalization level as SGD. However, the distributional dynamics of gRDA is drastically different from that of SGD. Specifically, an autoregressive soft-thresholding operator enters the distribution dynamics of gRDA, which encourages sparsity. Theoretical insights are provided to guide the selection of hyperparameters, which is validated by empirical analysis using CIFAR-10.

Single Index Models for Analysis of Mental Health Data with Functional Covariates

Debajyoti Sinha
FLORIDA STATE UNIVERSITY
E-mail: sinhad@stat.fsu.edu

Abstract: Single-index models are practical, useful tools for modeling and analyzing many clinical and psychological studies with complex non-linear covariate effects on the response. We propose frequentist and Bayesian methods for monotone single-index models where the monotonicity of the unknown link function renders a clinically interpretable index, along with the relative importance of the scalar and functional covariates on the index. To ease the computational complexity of the frequentist and Bayesian analysis, we also develop a novel and efficient algorithms. These methodologies and their advantages over existing methods are illustrated via simulation studies and analysis of a depression study of adolescent girls.

Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression

Liping Zhu
Renmin University of China
E-mail: zhu.liping@ruc.edu.cn

Abstract: Quadratic regression goes beyond the linear model by simu ltaneously including main effects and interactions between the covariat es. The problem of interaction estimation in high dimensional quadrati c regression has received extensive attention in the past decade. In th is article we introduce a novel method which allows us to estimate t

he main effects and interactions separately. Unlike existing methods f or ultrahigh dimensional quadratic regressions, our proposal does not r equire the widely used heredity assumption. In addition, our proposed estimates have explicit formulas and obey the invariance principle at the population level. We estimate the interactions of matrix form un der penalized convex loss function. The resulting estimates are shown to be consistent even when the covariate dimension is an exponentia l order of the sample size. We develop an efficient ADMM algorithm to implement the penalized estimation. This ADMM algorithm fully explores the cheap computational cost of matrix multiplication and is much more efficient than existing penalized methods such as the all-p airs
LASSO. We demonstrate the promising performance of our proposal t hrough extensive numerical studies.

S119: Recent Advances in High dimensional Statistics

Limit distribution theory in multiple isotonic regression

Cun-Hui Zhang
Rutgers University
E-mail: czhang@stat.rutgers.edu

Abstract: We study limit distributions for the tuning-free max-min block estimators in multiple isotonic regression under both fixed lattice design and random design settings. We show that at a fixed interior point in the design space, the estimation error of the max-min block estimator converges in distribution to a non-Gaussian limit at certain rate depending on the number of vanishing derivatives and certain effective dimension and sample size that drive the asymptotic theory. The limiting distribution can be viewed as a generalizing the well-known Chernoff distribution in univariate problems. The convergence rate is optimal in a local asymptotic minimax sense. There are two interesting features in our local theory. First, the max-min block estimator automatically adapts to the full spectrum of local smoothness levels and the intrinsic dimension of the isotonic regression function at the optimal rate. Second, the optimally adaptive local rates are in general not the same in fixed lattice and random designs. In fact, the local rate in the fixed lattice design case is no slower than that in the random design case, and can be much faster when the local smoothness levels of the isotonic regression function or the sizes of the lattice differ substantially along different dimensions. This is joint work with Qiyang Han.

Bayesian variance estimation in the Gaussian sequence model with partial information on the means

Gianluca Finocchio
University of Twente
E-mail: g.finocchio@utwente.nl

Abstract: Consider the Gaussian sequence model under the additional assumption that a fraction of the means are known. We study the problem of variance estimation from a frequentist Bayesian perspective. The maximum likelihood estimator (MLE) for the variance is biased and inconsistent. This raises the question whether the posterior is able to correct the MLE in this case. By developing a new proving strategy that uses refined properties of the posterior distribution, we find that the marginal posterior is inconsistent for any i.i.d. prior on the mean parameters. In particular, no assumption on the decay of the prior needs to be imposed. Surprisingly, we also find that consistency can be retained for a hierarchical prior based on Gaussian mixtures. In this case we also establish a limiting shape result and determine the limit distribution. In contrast to the classical Bernstein-von Mises theorem, the limit is non-Gaussian. By conducting a small numerical study, we show that the Bayesian analysis leads then to new statistical estimators outperforming the correctly calibrated MLE in a numerical simulation study.

Large Covariance Regression for Spatial Data

Wei Lin
Peking University
E-mail: weilin@math.pku.edu.cn

Abstract: Data assimilation plays an important role in geoscientific research, which can be viewed as inference in nonlinear, non-Gaussian state-space models with very high-dimensional state vectors. A Monte Carlo variant of the Kalman filter, known as the ensemble Kalman filter (EnKF), has been extremely successful in such applications but far from being well understood. A mystery in understanding the behavior of EnKF is the covariance inflation and localization that is often required in the EnKF updates. We cast this problem in the framework of large covariance regression for spatial data, and investigate the consistency and optimality of several estimators for the inflation factors. The methodology and theory are illustrated on simulated and real air pollution data.

The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy

Linjun Zhang
Rutgers University
E-mail: zlj11112222@gmail.com

Abstract: Privacy-preserving data analysis is a rising challenge in contem- porary statistics, as the privacy guarantees of statistical methods are often achieved at the expense of accuracy. In this paper, we investigate the tradeoff between statistical accuracy and privacy in mean estimation and linear regression, under both the classical low- dimensional and modern high-dimensional settings. A primary focus is to establish minimax optimality for statistical estimation with the (ε, δ)-differential privacy constraint. To this end, we find that classical lower bound arguments fail to yield sharp results, and new technical tools are called for.
Inspired by the theoretical computer science literature on “trac- ing adversaries”, we formulate a general lower bound argument for minimax risks with differential privacy constraints, and apply this argument to high-dimensional mean estimation and linear regression problems. We also design computationally efficient algorithms that attain the minimax lower bounds up to a logarithmic factor. In par- ticular, for the high-dimensional linear regression, a novel private iter- ative hard thresholding pursuit algorithm is proposed, based on a pri- vately truncated version of stochastic gradient descent. The numerical performance of these algorithms is demonstrated by simulation stud- ies and applications to real data containing sensitive information, for which privacy-preserving statistical methods are necessary.

S120: High dimensional Statistics and Probability

Refined Cramer type moderate deviation thorems for general self-nomalied sums with applications

Qi-Man Shao
Southern University of Science and Technology
E-mail: shaoqm@sustech.edu.cn

Abstract: "Let (X_i, Y_i) $1 leq i leq n$ be a sequence of independent random vectors. A refined Cramer type moderate deviation theorem for the self-normalized sum $(sum_{i=1}^n X_i) /(sum_{i=1}^n

Y_i^2)^{1/2}$ is obtained. The result extends earlier results by Jing, Shao and Wang (2003) and Wang (2011). Application to dependent random variables, Huber's estimator and square-root LASSO will be discussed."

Tests for principal eigenvalues and eigenvectors

Xinghua Zheng
HKUST
E-mail: xhzheng@ust.hk

Abstract: We establish CLTs for the principal eigenvalues and eigenvectors under a large factor model setting. As an application, we develop two-sample tests for both the principal eigenvalues and principal eigenvectors, which can be used to detect structural breaks in large factor models. While there exist such tests, they can not distinguish between individual eigenvalues and/or eigenvectors. Our tests provide unique insights into the source of structural breaks.

Based on joint work with Jianqing Fan, Yingying Li and Ningning Xia.

Single eigenvalue fluctuations of sparse Erdős-Rényi graphs

Yukun He
University of Zurich
E-mail: yukun.he@math.uzh.ch

Abstract: We will first review some universality results for Wigner matrices and sparse random graphs, and then talk about some recent development on the fluctuation of individial eigenvalues of these models. In particular, we show that the bulk eigenvalues of sparse matrices fluctuate on a scale different from Wigner matrices, i.e. they exhibit a non-universal behaviour.

Asymptotic Normality of the Maximum Likelihood Estimators in ANOVA Models

Fengnan Gao
Fudan University
E-mail: fngao@fudan.edu.cn

Abstract: We consider the maximum likelihood estimators in the ANOVA models with increasing dimensions. A clean sufficient condition and a clear and simple proof are presented for its asymptotic normality. The proof relies on formulating the problem as the root finding problem with Newton-Raphson algorithms being applied to the score function. An excursion may be made to include more general cases with as little conditions as possible imposed on the design for the asymptotic normality to hold.

S121: Complex data analysis and its applications

Joint modeling of multivariate continuous and time-to-event data

Xinyuan Song
Chinese University of Hong Kong
E-mail: xysong@sta.cuhk.edu.hk

Abstract: We propose a joint modeling approach to jointly analyze multivariate continuous and time-to-event data. The proposed model is composed of three parts. The first part is an exploratory factor analysis model that summarizes latent factors through multivariate continuous 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 full Bayesian approach coupled with efficient 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 that 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.

Statistical learning for individualized asset allocation

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

Abstract: We establish a statistical learning framework for individualized asset allocation. A high-dimensional Q-learning methodology is proposed for continuous decision making. The proposed methodology enjoys desirable theoretical properties and facilitates valid statistical inference for optimal values. Empirically, the proposed statistical learning framework is exercised with Health and Retirement Study data. The results show that our proposed optimal individualized strategy improves individual financial well-being and surpasses benchmark strategies under a consumption-based utility framework.

Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach

Sai Li
University of Pennsylvania
E-mail: sai.li@pennmedicine.upenn.edu

Abstract: Linear mixed-effects models are widely used in analyzing clustered or repeated measures data.

We propose a quasi-likelihood approach for estimation and inference of the unknown parameters in linear mixed-effects models with high-dimensional fixed effects. The proposed method is applicable to general settings where the cluster sizes are possibly large or unbalanced. Regarding the fixed effects, we provide rate optimal estimators and valid inference procedures that are free of the assumptions on the specific structure of the variance components. Separately, rate optimal estimators of the variance components are derived under mild conditions. We prove that, under proper conditions, the convergence rate for estimating the variance components of the random effects does not depend on the accuracy of fixed effects estimation. Computationally, the algorithm involves convex optimization and is loop-free. The proposed method is assessed in various simulation settings and is applied to a real study regarding the associations between the body weight index and polymorphic markers in a heterogeneous stock mice population.

Generalized integration model for Improved Statistical Inference by Leveraging External Summary Data

Kai Yu
National Cancer Institute
E-mail: yuka@mail.nih.gov

Abstract: Meta-analysis has become a powerful tool for enhanced inference by gathering evidence from multiple sources. It pools summary-level data from different studies to improve estimating efficiency with the assumption that all participating studies are analyzed under the same statistical model. It is challenging to integrate external summary data calculated from different models with a newly conducted internal study in which individual-level data is collected. We develop a novel statistical

inference framework based on a novel generalized integration model, which effectively synthesizes internal and external information for integrative analysis. The new framework is versatile enough to incorporate various types of summary data from multiple sources. We establish asymptotic properties for the proposed procedure and prove that the new estimate is theoretically more efficient than the internal data based maximum likelihood estimate, as well as a recently developed constrained maximum likelihood approach that incorporates the outside information. We illustrate an application of our method by evaluating cervical cancer risk using data from a large cervical screening program.

S122: Traditional statistical techniques in new data setting

Modeling Count Time Series via Common Factors

Fangfang Wang
Worcester Polytechnic Institute
E-mail: fwang4@wpi.edu

Abstract: In this talk, a new parameter-driven model for multivariate time series of counts is discussed. The time series is not necessarily stationary. We model the mean process as the product of modulating factors and unobserved stationary processes. The former characterizes the long-run movement in the data, while the latter is responsible for rapid fluctuations and other unknown or unavailable covariates. The unobserved stationary processes evolve independently of the past observed counts, and might interact with each other. We express the multivariate unobserved stationary processes as a linear combination of possibly low-dimensional factors that govern the contemporaneous and serial correlation within and across the observed counts. Regression coefficients in the modulating factors are estimated via pseudo maximum likelihood estimation, and identification of common factor(s) is carried out through eigenanalysis on a positive definite matrix that pertains to the autocovariance of the observed counts at nonzero lags. Theoretical validity of the two-step estimation procedure is presented. We also provide numerical and empirical results that corroborate the theoretical findings.

Variable Selection for Multiple Types of High-Dimensional Features With Missing Data

Kin Yau Wong
The Hong Kong Polytechnic University
E-mail: kin-yau.wong@polyu.edu.hk

Abstract: Recent technological advances have made it possible to collect multiple types of high-dimensional data in biological, clinical, and epidemiological studies. However, some data types or features may not be measured for all study subjects because of cost or other constraints. A common strategy for handling incomplete (high-dimensional) data is to obtain a complete data set using listwise deletion or through single imputation and then apply conventional statistical methods to the complete data set. This two-step approach, however, is inefficient and may even be biased. In this presentation, we present a valid and efficient approach to variable selection with multiple types of potentially missing features. We use a latent variable model to characterize the relationships across and within data types and to infer missing values from observed data. We develop a penalized-likelihood approach for variable selection and parameter estimation and devise an efficient expectation-maximization (EM) algorithm to implement our approach. The likelihood-based framework accommodates general missing-data patterns, and the low-dimensional factor model makes the estimation computationally tractable. We provide an application to a motivating multi-platform genomics study.

High-order Imaging Regression via Internal Variation

Long Feng
City University of Hong Kong
E-mail: longfeng@cityu.edu.hk

Abstract: The use of brain-imaging data to analyze cognitive disabilities has drawn increasing attentions in both psychology and public health. As brain-imaging data are usually represented by three or even higher order tensors, we aim to develop a class of high-order tensor (imaging) regression models to address this issue. A key novelty of our method is that it is able to account for the piecewise smooth nature of most imaging coefficients in the form of high-order tensors. This is achieved by an innovative approach named SHrinkage via Internal Variation (SHIV). The Internal Variation (IV) is designed to serve as a substitute of total variation (TV) for high order tensors. The SHIV is an IV penalized estimation and can be solved easily by a sequence of generalized Lasso problem. Theoretically, we simultaneously provide the computational and statistical errors of the SHIV estimates under a restricted eigenvalue condition and certain initialization requirements. Numerically, we conducted two simulation studies to demonstrate the accuracy of our method in brain region identification and tensor estimation. Furthermore, we analyzed the Philadelphia Neurodevelopmental Cohort dataset which includes pre-processing magnetic resonance images. Our analysis identified a subregion of cingulate cortex as being associated with verbal reasoning ability.

Distributed Learning with Minimum Error Entropy Principle

Xin Guo
The Hong Kong Polytechnic University
E-mail: xinguo.math@gmail.com

Abstract: Minimum Error Entropy (MEE) principle is an important approach in Information Theoretical Learning (ITL). It is widely applied and studied in various fields for its robustness to noise. In this paper, we study a reproducing kernel-based distributed MEE algorithm, DMEE, which is designed to work with both fully supervised data and semi-supervised data. With fully supervised data, our proved learning rates equal the minimax optimal learning rates of the classical pointwise kernel-based regressions. Under the semi-supervised learning scenarios, we show that DMEE exploits unlabeled data effectively, in the sense that first, under the settings with weaker regularity assumptions, additional unlabeled data significantly improves the learning rates of DMEE. Second, with sufficient unlabeled data, labeled data can be distributed to many more computing nodes, that each node takes only O(1) labels, without spoiling the learning rates in terms of the number of labels.

S123: Big Data and Artificial Intelligence in Medicine: a Bright Future

Generating Real World Evidence for Non-Communicable Disease Control

Jim Li
Pfizer Inc
E-mail: Jim.Li@pfizer.com

Abstract: Real world data (RWD) can be claims and transactions for healthcare resource utilization, electronic health records, surveys, linked datasets and other digital data collected outside a traditional clinical trial. Non-communicable diseases (NCDs), including cardiovascular diseases,

cancers, respiratory diseases, and diabetes, tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behaviors factors. To prevent and control NCDs, it is important to understand the risk factors for disease progression and treatment patterns including treatment adherence. Because of the chronic nature of NCDs, real world evidence R(RWE), generated from RWD, about the risk factors and treatment patterns are particularly useful for gaining valuable insights for the prevention and control of NCDs. Methodology and examples will be discussed in this presentation.

Evaluation of the three-in-one team-based care model on hierarchical diagnosis and treatment patterns among patients with diabetes: a retrospective cohort study using Xiamen's regional electronic health records

Yuji Feng
Beijing Innomed Health and Medical Research Center
E-mail: fengyuji@hotmail.com

Abstract: Background: Xiamen is a pilot city in China for hierarchical diagnosis and treatment reform of non-communicable diseases, especially diabetes. Since 2012, Xiamen has implemented a program called the "three-in-one", a team-based care model for the treatment of diabetes, which involves collaboration between diabetes specialists, general practitioners, and health managers. In addition, the program provides financial incentives to improve care, as greater accessibility to medications through community health care centers (CHCs). The aim of this study was to evaluate the effectiveness of these policies in shifting visits from general hospitals to CHCs for the treatment of type 2 diabetes mellitus (T2DM).

Method and materials: A retrospective observational cohort study was conducted using Xiamen's regional electronic health record (EHR) database, which included 90% of all patients registered since 2012. Logistic regression was used to derive the adjusted odds ratio (OR) for patients shifting from general hospitals to CHCs. Among patients treated at hospitals, Kaplan-Meier(KM) curves were constructed to evaluate the time from each policy introduction until the switch to CHCs. A k-means clustering analysis was conducted to identify patterns of patient care-seeking behavior.

Results: In total, 89,558 patients and 2,373,524 visits were included. In contrast to increased outpatient visits to general hospitals in China overall, the percentage of visits to CHCs in Xiamen increased from 29.7% in 2012 to 66.5% in 2016. The most significant and rapid shift occurred in later periods after full policy implementation. Three clusters of patients were identified with different levels of complications and health care-seeking frequency. All had similar responses to the policies.

Conclusions: The "three-in-one" team-based care model showed promising results for building a hierarchical health care system in China. These policy reforms effectively increased CHCs utilization among diabetic patients.

Keywords: Health policy reform, Chronic disease, Policy evaluation, Hierarchical health care

Machine Learning and Artificial Intelligence for Healthcare

Haoda Fu
Eli Lilly and Company
E-mail: fu_haoda@lilly.com

Abstract: in this talk we will share our journey on digital health using machine learning and AI technology. In particular, we will focus on 4 areas including automatic control, deep learning, recommendation system, and reinforcement learning. We will also share experience and examples on the common mistakes that inexperience data scientists are made. It will make a point that statistics are still essential in applying machine learning models.

ShortBio: Dr. Haoda Fu is a senior research advisor and a enterprise lead for Machine Learning, Artificial Intelligence, and Digital Connected Care from Eli Lilly and Company. Dr. Haoda Fu is a Fellow of ASA (American Statistical Association). He is also an adjunct professor of biostatistics department, Indiana university school of medicine. Dr. Fu received his Ph.D. in statistics from University of Wisconsin - Madison in 2007 and joined Lilly after that. Since he joined Lilly, he is very active in statistics methodology research. He has more than 90 publications in the areas, such as Bayesian adaptive design, survival analysis, recurrent event modeling, personalized medicine, indirect and mixed treatment comparison, joint modeling, Bayesian decision making, and rare events analysis. In recent years, his research area focuses on machine learning and artificial intelligence. His research has been published in various top journals including JASA, JRSS, Biometrics, ACM, IEEE, JAMA, Annals of Internal Medicine etc.. He has been teaching topics of machine learning and AI in large industry conferences including teaching this topic in FDA workshop. He was board of directors for statistics organizations and program chairs, committee chairs such as ICSA, ENAR, and ASA Biopharm session.

S124: Novel approaches for analysis of probability and non-probability samples

Combining Probability Non-probability Samples: Theory and Practice

Michael Elliott
University of Michigan
J-mail: mrelliot@umich.edu

Abstract: Although probability sample designs remain a "gold standard" in survey research, demand for use of non-probability samples is increasing, due to, among other reasons, rising costs and falling response rates in probability samples and the availability of "big data" from administrative databases, social media users, and other sources. Design-based inference, in which the distribution for inference is generated by the random mechanism used by the sampler, cannot be used for non-probability samples. One alternative is quasi-randomization in which pseudo-inclusion probabilities are estimated based on covariates available for samples and nonsample units. Another is superpopulation modeling for the analytic variables collected on the sample units in which the model is used to predict values for the nonsample units. A third alternative is a model-assisted approach in which probability samples are used to develop calibration estimators. We will overview these approaches and discuss their unique advantages in different analytic and application settings.

Hypotheses Testing from Complex Survey Data Using Bootstrap Weights: A Unified Approach

Zhonglei Wang
Xiamen University
E-mail: wangzl@xmu.edu.cn

Abstract: Standard statistical methods that do not take proper account of the complexity of survey design can lead to erroneous inferences when applied to survey data due to unequal selection probabilities, clustering, and other design features. In particular, the actual type I error rates of tests of hypotheses using standard methods can be much bigger than the nominal

significance level. Methods that take account of survey design features in testing hypotheses have been proposed, including Wald tests and quasi-score tests that involve the estimated covariance matrices of parameter estimates. In this paper, we present a unified approach to hypothesis testing that does not require computing the covariance matrices by constructing bootstrap approximations to weighted likelihood ratio statistics and weighted quasi-score statistics and establish the asymptotic validity of the proposed bootstrap tests. In addition, we also consider hypothesis testing from categorical data and present a bootstrap procedure for testing simple goodness of fit and independence in a two-way table. In the simulation studies, the type I error rates of the proposed approach are much closer to their nominal level compared with the naive likelihood-ratio test and quasi-score test. An application to data from an educational survey under a logistic regression model is also presented.

Methodologies for Analyzing Non-probability Survey Samples

Changbao Wu
University of Waterloo
E-mail: cbwu@uwaterloo.ca

Abstract: We provide an overview on recent developments for analyzing non-probability survey samples. Inferential frameworks and theoretical results on sample matching and double robust estimation are discussed, and finite sample performances of the estimators are examined through simulation studies. An application to analyzing a non-probability survey sample from the PEW Research Centre is presented. Some practical issues are also discussed.

Bayesian Inference for Sample Surveys in the Presence of High-Dimensional Auxiliary Information

Qixuan Chen
Columbia University
E-mail: qc2138@cumc.columbia.edu

Abstract: Survey inference can be challenged by non-representativeness of survey samples, either imperfect probability samples or non-probability samples without a probability sampling design. We consider improving survey inference with a non-representative survey sample in the presence of high-dimensional auxiliary information, which are measured in the survey sample and also available about the population via such as census data or administrative records. We propose Bayesian model-based predictive methods for estimating finite population totals by modeling the conditional distribution of the survey outcome using Bayesian additive regression trees (BARTs), which naturally handles high-dimensional auxiliary variables allowing possible interactions and nonlinear associations. Besides the auxiliary variables, inspired by Little and An (2004), we estimate the propensity score for a unit to be included in the sample using another BART. We include both the propensity score and key predictors of the survey outcome as covariates in the model to improve the estimation of population totals. We show through simulations studies and a real survey that the Bayesian model-based methods using BARTs improve survey inference.

S125: Advances in Statistical Analysis of Omics Data in Agriculture

An Efficient Statistical Method for Genomic Selection

Min Zhang
Purdue University
E-mail: minzhang@stat.purdue.edu

Abstract: The high-throughput genotyping technology has generated a huge number of genomic markers that can be used for genomic selection. However, the large number of markers makes it difficult to estimate the breeding values. We propose to apply the penalized orthogonal-components regression method to estimate breeding values. As a supervised dimension reduction method, it can sequentially constructs linear combinations of markers, i.e. orthogonal components, such that these components are closely correlated to the phenotype. Such a dimension reduction is able to group highly correlated predictors and allows for collinear or nearly collinear markers. As shown in simulation studies, the proposed method is computationally efficient and provides accurate prediction when compared to existing methods. IN addition, the utility of the method was demonstrated through applications to real data.

Statistics Improves Effectiveness of Genomic Selection in Plant Breeding

Lan Zhu
Oklahoma State University
E-mail: lan.zhu@okstate.edu

Abstract: The genomic revolution opened up the possibility for predicting un-tested phenotypes in schemes commonly referred as genomic selection (GS). Considering the practicality of applying GS in the line development stage of a hard red winter (HRW) wheat variety development program (VDP), effectiveness of GS was evaluated by prediction accuracy, as well as by the response to selection across field seasons that demonstrated challenges for crop improvement under significant climate variability. Important breeding targets for HRW wheat improvement in the southern Great Plains of USA, including Grain Yield, Kernel Weight, Wheat Protein content, and Sodium Dodecyl Sulfate (SDS) Sedimentation Volume as a rapid test for predicting breadmaking quality, were used to estimate GS's effectiveness across harvest years from 2014 (drought) to 2016 (normal). In general, nonparametric algorithms RKHS and RF produced higher accuracies in both same-year/environment cross validations and cross-year/environment predictions, for the purpose of line selection in this bi-parental doubled haploid (DH) population. Further, the stability of GS performance was greatest for SDS Sedimentation Volume but least for Wheat Protein content. To ensure long-term genetic gain, our study on selection response suggested that across this sample of environmental variability, and though there are cases where phenotypic selection (PS) might be still preferential, training conducted under drought stress or in suboptimal conditions could still provide an encouraging prediction outcome, when selection decisions were made in normal conditions. However, it is not advisable to use training information collected from a normal field season to predict trait performance under drought conditions. Further, the superiority of response to selection was most evident if the training population can be optimized.

Exploring high-throughput plant phenomics and genomics data

Yumou Qiu
Iowa State University
E-mail: yumouqiu@iastate.edu

Abstract: High-throughput phenotyping systems provide abundant data for statistical analysis through plant imaging. Before usable data can be obtained, image processing must take place. Unlike well-established pipelines for processing genomics data, the analysis of phenomics data is a current bottleneck for Omics studies. We propose the use of supervised learning methods to segment plants from background in plant images and

compares them to commonly used thresholding methods. As obtaining accurate training data is a major obstacle to using supervised learning methods for segmentation, a novel approach to producing accurate labels is proposed. It is demonstrated that with careful selection of training data through such an approach, supervised learning methods, and neural networks in particular, can outperform thresholding methods at segmentation.

Feature Selection for Rhizosphere Microbiome Studies in Presence of Confounding Using Standardization

Peng Liu
Iowa State University
E-mail: pliu@iastate.edu

Abstract: Microbiome studies have uncovered associations between microbes and plant, animal, and human health outcomes. This has led to an interest in identifying microbial interventions for treatment of disease and optimization of crop yields which will require the identification of individual relevant microbiome features. That task is challenging because of the high dimensionality of microbiome data and the confounding that results from the complex and dynamic interactions among host, environment, and microbiome. The performance of variable selection and estimation procedures may be unsatisfactory when there are differentially abundant features resulting from a categorical confounding variable. For microbiome studies with such a confounding structure, we propose a standardization approach to estimation of population effects of individual microbiome features. Due to the high dimensionality and confounding-induced correlation between features, we propose feature screening, selection, and estimation conditional on each stratum of the confounder. Comprehensive simulation studies are used to demonstrate the advantages of our approach in recovering relevant features. Utilizing a potential-outcomes framework, we outline assumptions required to ascribe causal, rather than associational, interpretations to the identified microbiome effects. We conducted an agricultural study of the rhizosphere microbiome of sorghum in which nitrogen fertilizer application is a confounding variable. We applied the proposed approach and identified microbial taxa that are consistent with biological understanding of potential plant-microbe interactions.

S126: Novel Bayesian Adaptive Clinical Trial Designs for Immunotherapy and Precision Medicine

ComPAS: A Novel Bayesian drug combination platform trial design with adaptive shrinkage for I/O check inhibitors

Sammi Tang
Servier Pharmaceuticals
E-mail: Sammi.tang@servier.com

Abstract: Combining different treatment regimens provides an effective approach to induce a synergistic treatment effect and overcome resistance to monotherapy. It will be the trend for I/O check inhibitors too. The challenge is that, given the large number of existing therapies, the number of possible combinations is huge and new potentially more efficacious compounds may become available any time during drug development.

To address this challenge, we proposed a flexible Bayesian drug combination platform design with adaptive shrinkage (ComPAS), which allows for dropping futile combinations, graduating effective combinations, and adding new combinations during the course of the trial. A new adaptive shrinkage method is developed to adaptively borrow information across combinations and efficiently identify the efficacious combinations based on Bayesian model selection and hierarchical models. Simulation studies show that ComPAS identifies the effective

combinations with higher probability than some existing designs. ComPAS provides an efficient and flexible platform to accelerate drug development in a seamless and timely fashion. Paper is published in the journal of Statistics in Medicine and interactive R shiny app is available to public.

Incorporating population pharmacokinetics data for Phase I-II dose-schedule finding

Fangrong Yan
China Pharmaceutical University
E-mail: f.r.yan@163.com

Abstract: In early phase clinical trials, incorporating PK information into dose finding process is an important thought. PK information could be considered as an appropriate indicator for evaluating the degree of drug intervention in humans. FDA has issued the population pharmacokinetics about the guidance for industry to guide the population analysis since it could be used to guide drug development and provide guidance about dose individualization. However, traditional clinical trial designs usually execute dose finding and PK analysis separately while the object of general dose finding is to identify an optimal dose of a treatment with a fixed schedule, which may cause over or under exposure of patients. Thus, we propose a novel Phase I-II dose-schedule finding design incorporating PK information to improve the design by taking advantage of the PK information collected from patients as well. The AUC indicator calculated from population pharmacokinetics is considered into the model to combine PK analysis with the probability models of both toxicity and efficacy, and construct the joint effects though utility function for combination allocation. Simulations illustrates that the design we proposed has a good ability of making the correct dose-schedule combination selection.

Characteristics of early phase trial designs for immunocology and comparisons of common designs

Yaqian Zhu
University of Pennsylvania
E-mail: yazhu@pennmedicine.upenn.edu

Abstract: Many new developments in cancer treatment in the modern era involve using patients' own immune system to fight cancer. Therapeutic options under this framework such as immune checkpoint inhibitors and monoclonal antibodies are all success examples of immunotherapy. These agents have demonstrated promising clinical activity across many disease indications but also present new challenges in the design and analysis of the early phase clinical trials for those agents. In this presentation, we will describe several unique characteristics of these therapy options including different toxicity profiles and mechanisms of action for which the classic statistical assumptions typically associated for cytotoxic agents may no longer be applicable. We will also review a few popular designs in the literature, including the 3+3, continuous reassessment method (CRM), Bayesian optimal interval (BOIN) design, and Keyboard design, and evaluate how varying design parameters, such as number of dose levels, target toxicity rate, maximum sample size, and cohort size, could impact the performances of each design through simulations. We will focus on parameter specifications that are commonly used in real world clinical trials of immunotherapy agents. Our preliminary results indicate that 3+3 tends to

have the worst performance while BOIN and Keyboard perform similarly to the CRM.

S127: Advances in Large Scale Data Analysis

Bayesian Analysis of Multidimensional Functional Data

Donatello Telesca
UCLA
E-mail: dtelesca@ucla.edu

Abstract: Multi-dimensional functional data arises in numerous modern scientific experimental and ob- servational studies. In this lecture we focus on longitudinal functional data, a structured form of multidimensional functional data. Operating within a longitudinal functional framework we aim to capture low dimensional interpretable features. We propose a computationally efficient nonparametric Bayesian method to simultaneously smooth observed data, estimate conditional functional means and functional covariance surfaces. Statistical inference is based on Monte Carlo samples from the posterior measure through adaptive blocked Gibbs sampling. Several operative characteristics associated with the proposed modeling framework are assessed comparatively in a simulated environment. We illustrate the application of our work in two case studies. The first case study involves age-specific fertility collected over time for various countries. The second case study is an implicit learning experiment in children with Autism Spectrum Disorder (ASD).

A New Joint Screening Method for Right-Censored Time-to-Event Data with Ultra-high Dimensional Covariates

Yi Liu
Ocean University of China
E-mail: liuyi@amss.ac.cn

Abstract: In an ultra-high dimensional setting with a huge number of covariates, variable screening is useful for dimension reduction before applying a more refined method for model selection and statistical analysis. This paper proposes a new sure joint screening procedure for right-censored time-to-event data based on a sparsity-restricted semiparametric accelerated failure time model. Our method, referred to as Buckley-James assisted sure screening (BJASS), consists of an initial screening step using a sparsity-restricted least-squares estimate based on a synthetic time variable and a refinement screening step using a sparsity-restricted least-squares estimate with the Buckley-James imputed event times. The refinement step may be repeated several times to obtain more stable results. We show that with any fixed number of refinement steps, the BJASS procedure retains all important variables with probability tending to 1. Simulation results are presented to illustrate its performance in comparison with some marginal screening methods. Real data examples are provided using a diffuse large-B-cell lymphoma (DLBCL) data and a breast cancer data. We have implemented the BJASS method using Matlab and made it available to readers through Github https://github.com/yiucla/BJASS.

Simultaneous estimation and variable selection for Interval-Censored Data with Broken Adaptive Ridge Regression

Hui Zhao
Zhongnan University of Economics and Law
E-mail: hzhao@mail.ccnu.edu.cn

Abstract: The simultaneous estimation and variable selection for Cox model has been discussed by several authors (Fan and Li, 2002; Huang and Ma, 2010; Tibshirani, 1997) when one observes right-censored failure time data. However, there does not seem to exist an established procedure for interval-censored data, a more general and complex type of failure time data, except two parametric procedures in Scolas et al. (2016) and Wu and Cook (2015). To address this, we propose a broken adaptive ridge (BAR) regression procedure that combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage. In particular, the method allows for the number of covariates to be diverging with the sample size. Under some weak regularity conditions, unlike most of the existing variable selection methods, we establish both the oracle property and the grouping effect of the proposed BAR procedure. We conduct an extensive simulation study and show that the proposed approach works well in practical situations and deals with the collinearity problem better than the other oracle-like methods. An application is also provided.

Dimension Reduction via cross-validation metric learning

Linlin DAI
Southwestern University of Finance and Economics
E-mail: ldaiab@swufe.edu.cn

Abstract: We propose a cross-validation metric learning approach to l earn a distance metric for dimension reduction in the multiple-index model. We minimize a leave-one-out cross-validation-type loss functio n, where the unknown link function is approximated by a metric-base d kernel-smoothing function. In contrast to existing methods, the new method requires very weak assumption on the design of predictors a nd is relatively easy-to-implement numerically.

S128: Real world evidence in medical research: methods and applications

Targeted integrative learning with applications in suicide risk prediction

Kun Chen
University of Connecticut
E-mail: kun.chen@uconn.edu

Abstract: In many scientific problems, the goal is to make inference on a specified "target population" of interest. For example, in a single-arm clinical trial, the target population can be defined by the treated patients and the key is to find out what happens to them if they were not treated; in a suicide risk study, the target population may be consist of patients who received care from a specific healthcare provider. Yet, the data available may go way beyond the target population. As such, a crucial question is how to best integrate all the information to improve the inference for the target. In this talk, we consider two scenarios. For the scenario of "integrated data", we propose a distance-segmented regression (DSR) framework, in which a distance metric is used to measure how close each sample is to the target and is assumed to guide the conditional association between the outcome and predictors. For the scenario of "non-integratable" data, we propose a transfer learning model, in which the target population and the external database are linked through subject similarities. Applications in suicide risk prediction with medical claim data will be discussed.

Analysis of semi-competing risks data using Archimedean copula models

Antai Wang
New Jersey Institute of Technology
E-mail: antai.wang@njit.edu

Abstract: In this talk, we propose a new method to analyze semi-competing

risks data using Archimedean copula models. Our method is simpler than the existing one and tends to be less biased. We illustrate our method using an example.

Perspective Plan for Studies Combining Real World Data Sources

Gang Li
Janssen
E-mail: ligang8844@gmail.com

Abstract: Many European countries have national registry databases. These data sources provide opportunities for medical research. However, for some research, the sample size in each data source is too small to draw a robust conclusion. But the privacy regulations in EU hinders researchers to combine several data sources together to conduct the analysis. In this research we will explore a 2-step analysis strategy: 1) the analysis is conducted on each data source that the consistent aggregated / summary data are generated for all sources; 2) the summary data from individual source will be combined for meta-analysis. In addition we will examine what are the additional information that is potentially useful for the meta-analysis, e.g., corrections of 2 variables.

Two-Stage Multi-Factor Adaptive Clinical Trials

Samuel Wu
University of Florida
E-mail: samwu@biostat.ufl.edu

Abstract: Efficacious and effective interventions usually involve multiple factors. And it is fantasy to have pilot information about the main effects and their interactions for all factors at the time of initial study design. Hence it will be beneficial for such studies to adopt adaptive designs that use continuously updated data to modify certain aspects of trial characteristics without undermining its validity. There has been increasing interest in adaptive trial designs that appropriately use accrued data over time to modify key trial characteristics in order to improve trial efficiency and gain ethical benefits. In this talk we will present some adaptive strategies that explore the specific structure of factorial designs based on two real studies. In addition, we will discuss new statistical design and inference procedures for two-stage multi-factor adaptive clinical trials to identify active factors.

S129: Integrative Analyses for Wearable Sensor Data in Clinical Studies

Dynamic Bayesian Prediction and Calibration using Multivariate Sensor Data Streams

Zhenke Wu
University of Michigan
E-mail: zhenkewu@umich.edu

Abstract: There is a critical need to understand the temporal dynamics of depression using real-time, objective measures. We introduce a flexible multivariate time series model to analyze multiple sensor data streams collected at distinct time scales (minutely, daily, and quarterly) with occasional missingness (due to failure to wear wristbands or carry smartphones). Our model predicts interns' mood and estimates the profile of lagged effects for each predictor time series by sharing information both across time, to account for smooth time-varying associations, as well as across similar subjects. We illustrate our methods using data from the 2017-18 Intern Health Study cohort recruited at University of Michigan. Lastly, we discuss computational issues and the practical implications of our results in the analysis of emerging intensive longitudinal data in mobile health.

Body Posture Recognition Based on the Raw Accelerometry Data

Jaroslaw Harezlak
Indiana University School of Public Health
E-mail: harezlak@iu.edu

Abstract: Applications of novel statistical and signal processing techniques to the raw accelerometry data enable reliable estimation of physical activity in a free-living environment. We present our work quantifying non-sedentary behavior based on the measurements from a tri-axial wrist-worn accelerometer. The key idea leverages the observation that hands are pointed down during standing and pointed mostly horizontally while sitting. In addition, standing activities often generate signal with large sub-second variation. Our algorithm, Sedentary and Upright Body Posture Classification (SedUp), utilizes data obtained from the axis representing the spatial position of a sensor. We estimate median and median standard deviation for the chosen axis in sliding time windows. The classification between sedentary and upright body posture is obtained using the logistic mixed regression models built using the extracted features. The method is applied to the data collected in a cohort of HIV-infected individuals who wore the devices for one week in a free-living environment. We summarize the associations between the health outcomes and extracted physical activity measures, such as amount of time standing per day.

Which to use: objective measurement or performance test for physical activity?

Jiawei Bai
Johns Hopkins University
E-mail: jbai@jhsph.edu

Abstract: Objective measurement of physical activity has become an important part of many studies that include assessment of human physical function. Accelerometer based wearable devices are the major tool utilized in such studies, because they can be worn on the human body relatively comfortably for an extended period of time 鈥?this enables a rich data collection in the free-living environment for weeks. However, some traditional methods such as performance tests are still used in many applications, because they are well-studied in the literature and often provide more detailed information on some specific function. In this paper, we introduce a series of statistical tools and models to assess and compare, on the same population, what information about the physical function we can get from free-living accelerometry measurement and in-lab performance tests. We used the data from the OUTLET Study of the METRC (Major Extremity Trauma Research Consortium), which aimed to compare 18-month functional outcomes and health related quality of life of patients undergoing salvage versus amputation following severe leg/foot injury.

Functional Marginal Structural Models for Time-varying Confounding of Mood Assessments

Haochang Shou
University of Pennsylvania
E-mail: hshou@pennmedicine.upenn.edu

Abstract: The increasingly dense assessments of multiple biosignals have provided opportunities for us to objectively track time-updated biological, yet also posed challenges for statistical analysis. For example, in two families of mood disorders from the National Institutes of Mental Health

and Lausanne-Geneva, the participants were evaluated on their minute-by-minute physical activity intensities continuously over two weeks using accelerometers. Meanwhile, they also answered questionnaires about their mood and behaviors through ecological momentary assessments (EMA) several times a day. While we are interested in the time-varying effects of mood (sadness, anxiety etc.) on endpoint events such as migraine, we are aware that physical activity might potentially influence mood at the next time point. Hence we propose a marginal structural models (MSM) for functional data such as continuous daily physical activity profiles and use inverse probability weighting to correct for potential bias induced by the time-dependent confounding effects of physical activities on mood.

S130: The Advances of Powerful Tools for Complex Neuroimaging Data

Learning Signal Subgraphs from Longitudinal Brain Networks with Symmetric Bilinear Logistic Regression

Lu Wang
Central South University
E-mail: wanglu_stat@csu.edu.cn

Abstract: Modern neuroimaging technologies, combined with state-of-the-art data processing pipelines, have made it possible to collect longitudinal observations of an individual's brain connectome at different ages. It is of substantial scientific interest to study how brain connectivity varies over time in relation to human cognitive traits. In brain connectomics, the structural brain network for an individual corresponds to a set of interconnections among brain regions. We propose a symmetric bilinear logistic regression to learn a set of small subgraphs relevant to a binary outcome from longitudinal brain networks as well as estimating the time effects of the subgraphs. We enforce the extracted signal subgraphs to have clique structure which has appealing interpretations as they can be related to neurological circuits. The time effect of each signal subgraph reflects how its predictive effect on the outcome varies over time, which may improve our understanding of interactions between the aging of brain structure and neurological disorders. Application of this method on longitudinal brain connectomics and cognitive capacity data shows interesting discovery of relevant interconnections among a small set of brain regions in frontal and temporal lobes with better predictive performance than competitors.

Classifying EEG Functional Connectivity Patterns Using A Multi-Domain Convolutional Neural Network

Chee-Ming Ting
Universiti Teknologi Malaysia
E-mail: cmting@utm.my

Abstract: We exploit altered patterns in brain functional connectivity as features for automatic discriminative analysis of neuropsychiatric patients. Deep learning methods have been introduced to functional network classification only very recently for fMRI, and the proposed architectures essentially focused on a single type of connectivity measure. We propose a deep convolutional neural network (CNN) framework for classification of electroencephalogram (EEG)-derived brain connectome in schizophrenia (SZ). To capture complementary aspects of disrupted connectivity in SZ, we explore combination of various connectivity features consisting of time and frequency-domain metrics of effective connectivity based on vector autoregressive model and partial directed coherence, and complex network measures of network topology. We design a novel multi-domain connectome CNN (MDC-CNN) based on a parallel ensemble of 1D and 2D CNNs to integrate the features from various domains and dimensions using different fusion strategies. We also consider an extension to dynamic brain connectivity using the recurrent neural networks. Hierarchical latent representations learned by the multiple convolutional layers from EEG connectivity reveals apparent group differences between SZ and healthy controls (HC). Evaluated on resting-state EEG data, the proposed MDC-CNN by integrating information from diverse brain connectivity descriptors is able to accurately discriminate SZ from HC, outperforming support vector machines.

A functional mixed model for scalar on function regression with application to a functional MRI study

Luo Xiao
North Carolina State University
E-mail: lxiao5@ncsu.edu

Abstract: Motivated by a functional magnetic resonance imaging (MRI) study, we propose a new functional mixed model for scalar on function regression. The model extends the standard scalar on function regression for repeated outcomes by incorporating subject-specific random functional effects. Using functional principal component analysis, the new model can be reformulated as a mixed effects model and thus easily fit. A test is also proposed to assess the existence of the subject- specific random functional effects. We evaluate the performance of the model and test via a simulation study, as well as on data from the motivating fMRI study of thermal pain. The data application indicates significant subject-specific effects of the human brain hemodynamics related to pain and provides insights on how the effects might differ across subjects.

Neuroconductor: An R Platform for Medical Imaging Analysis

John Muschelli
Johns Hopkins University
E-mail: muschellij2@gmail.com

Abstract: Neuroconductor (https://neuroconductor.org) is an open-source platform for rapid testing and dissemination of reproducible computational imaging software. The goals of the project are to: 1) provide a centralized repository of R software dedicated to image analysis, 2) disseminate software updates quickly, 3) train a large, diverse community of scientists using detailed tutorials and short courses, 4) increase software quality via automatic and manual quality controls, and 5) promote reproducibility of image data analysis.

We provide a description of the purpose of Neuroconductor, highlight some packages, and show some imaging analysis examples.

S131: Statistical Learning for the Analysis of Large-scale Omics Data

DeepHiC: Greatly Enhancing Chromatin Interaction Information Using Deep Learning

Chun Li
Case Western Reserve University
E-mail: lichun1668@gmail.com

Abstract: Hi-C is a high-throughput sequencing technique for studying chromatin conformations. Due to relatively high cost, a typical Hi-C dataset for a human sample has 150-300 million read pairs, which is often used for detecting chromatin conformations at the 40kb resolution. Results at finer resolutions require a lot more data (e.g., 10kb would require 3 billion read pairs). To enhance the quality of Hi-C data, we develop DeepHiC, a deep learning algorithm, to enhance the information of a typical dataset for

analysis at the 10kb resolution. We evaluate the method from several perspectives: Spearman's correlation, effective depth, and chromatin loop calling. We also demonstrate the performance of the algorithm for its potential use across tissue types and across species.

Measurement Errors in Array-Based DNA Methylation Analysis

Weihua Guan
University of Minnesota
E-mail: wguan@umn.edu

Abstract: Genome-wide DNA methylation measures are now routinely used to investigate their association with various outcomes of interest. Inherent to the array-based DNA methylation measures, e.g., Illumina HumanMethylation450 (HM450) or Infinium MethylationEPIC chips, is the associated technical variation of non-biological interest. One example of such variation is the bead-to-bead variation due to the use of multiple beads on HM450 or EPIC arrays. These technical variations can be viewed as measurement errors, and are commonly ignored in downstream association analysis. We have proposed a novel statistical framework to take into account of these measurement errors. Specifically, we used a mixed effects model to quantify the measurement error, and developed an expectation–maximization (EM) algorithm to estimate the model parameters. We applied our proposed method to the Atherosclerosis Risk in Communities (ARIC) methylation data (n = 2,843; HM450 array) in an epigenome-wide association study of smoking status, after accounting for the bead-to-bead variation. We identified 14 additional CpG sites associated with smoking, at sites with high technical variation (intraclass correlation coefficient < 0.4). We expect that our new method can improve statistical power of association tests and accuracy of parameter estimates in future epigenome-wide association studies (EWASs) using methylation arrays.

Statistical learning for analyzing single-cell multi-omics data

Wei Chen
University of Pittsburgh
E-mail: wec47@pitt.edu

Abstract: The rapid advances in single cell technologies have been playing important roles in understanding the heterogeneity and dynamics of various cell populations in complex multicellular tissue or organs. The recently developed droplet-based single cell transcriptome sequencing (scRNA-seq) technology enables researchers to measure the gene expression of tens of thousands single cells simultaneously. More recently, coupling with droplet-based scRNA-seq, another revolutionary technology named Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) allows the detection of cell surface proteins and transcriptome profiling within the same cell simultaneously. Despite the rapid advances in technologies, novel statistical methods and computation tools for analyzing single-cell multi-omics data are still lacking. In this study, we developed a novel random effects model that jointly analyze the paired data from scRNA-seq and CITE-seq experiments under a Bayesian framework. In the simulation study and analysis of in-house real data sets, we demonstrated the validity and advantages of our method in understanding immune cells as well as facilitating novel biological discoveries.

A powerful method for the estimation of cancer-driver genes using a weighted iterative zero-truncated negative-binomial regression

Miaoxin Li
Sun Yat-sen University
E-mail: limiaoxin@mail.sysu.edu.cn

Abstract: Genomic identification of driver mutations and genes in cancer cells are critical for precision medicine. Due to difficulty in modeling distribution of background mutations, existing statistical methods are often underpowered to discriminate driver genes from passenger genes. Here we propose a novel statistical approach, weighted iterative zero-truncated negative-binomial regression (WITER), to detect cancer-driver genes showing an excess of somatic mutations. By solving the problem of inaccurately modeling background mutations, this approach works even in small or moderate samples. Compared to alternative methods, it detected more significant and cancer-consensus genes in all tested cancers. Applying this approach, we estimated 178 driver genes in 26 different cancers types. In silico validation confirmed 90.5% of predicted genes as likely known drivers and 7 genes unique for individual cancers as very likely new drivers. The technical advances of WITER enable the detection of driver genes in TCGA datasets as small as 30 subjects, rescuing more genes missed by alternative tools. The tool is available at http://grass.cgs.hku.hk/limx/witer/ and http://grass.cgs.hku.hk/limx/kggseq/.

S132: Methods for measurement error problems and their role in improving EHR data-based discovery

Methods to address correlated exposure and outcome error for failure time outcomes

Pamela Shaw
Associate Professor of Biostatistics
E-mail: shawp@pennmedicine.upenn.edu

Abstract: Electronic health records (EHR) data are increasingly used in medical research, but these data are often subject to measurement error. These errors, if not addressed, can potentially bias results in association analyses. Methodology to address covariate measurement error has been well developed; however, methods to address errors in time-to-event outcomes are relatively underdeveloped. We will consider methods to address errors in both the covariate and time-to-event outcome that are potentially correlated. We develop an extension to the popular regression calibration method for this setting. Regression calibration has been shown to perform well for settings with covariate measurement error, but it is known that this method is generally biased for nonlinear regression models, such as the Cox model for time-to-event outcomes. Thus, we additionally propose raking estimators. Raking is a standard method in survey sampling that makes use of auxiliary information on the population to improve upon the simple Horvitz-Thompson estimator applied to a subset of data (e.g. the validation subset). Raking estimators are consistent when based on a consistent estimating equation for the validation subset. We demonstrate through numerical studies the relative performance of the regression calibration and raking estimators. We will discuss the choice of the auxiliary variable and aspects of the underlying estimation problem that affect the degree of improvement that the raking estimator will have over the simpler, biased regression calibration approach. We consider the relative performance under varying levels of signal, covariance, and censoring. We further illustrate the methods with a real data example using observational EHR data on HIV outcomes from the Vanderbilt Comprehensive Care Clinic.

Support Vector Machine with Measurement Error

Xin Liu

Shanghai University of Finance and Economics
E-mail: liu.xin@mail.shufe.edu.cn

Abstract: The Support Vector Machine (SVM) and its extensions have been widely used in various areas due to its great prediction capability. However, when the data are contaminated with measurement errors, the performance of an SVM may be deteriorated. In this research, we focus on the measurement error in the response variable with binary outcomes, and find that how it affects the classification accuracy. Correspondingly, we propose the response-error-corrected Support Vector Machine (REC-SVM) incorporating response measurement error into the standard SVM framework. Not only does the classifier outperform with error-contaminated response, but performs comparatively with error-free response as well. The excellent and robust performance of the response-error-corrected SVM is further demonstrated in the numeric investigations with synthetic data sets and a real prostate cancer data set.

Augmented methods in PheWAS studies for pleiotropic effects using biobank data

Yong Chen
University of Pennsylvania
E-mail: ychen123@upenn.edu

Abstract: Pleiotropy, where 1 genetic locus affects multiple phenotypes, can offer significant insights in understanding the complex genotype–phenotype relationship. Although individual genotype–phenotype association shave been thoroughly explored, seemingly unrelated phenotypes can be connected genetically through common pleiotropic loci or genes. However, current analyses of pleiotropy have been challenged by both methodologic limitations and a lack of available suitable data sources.

In this talk, we present a regression framework, reduced rank regression, to simultaneously analyze multiple phenotypes and genotypes to detect pleiotropic effects. We used a large-scale biobank linked electronic health record data from the Penn Medicine BioBank to select 5 cardiovascular diseases (hypertension, cardiac dysrhythmias, ischemic heart disease, congestive heart failure, and heart valve disorders) and 5 mental disorders (mood disorders; anxiety, phobic and dissociative disorders; alcohol related disorders; neurological disorders; and delirium dementia) to validate our framework.

Compared with existing methods, reduced rank regression showed a higher power to distinguish known associated single-nucleotide polymorphisms from random single-nucleotide polymorphisms. In addition, genome-wide gene-based investigation of pleiotropy showed that reduced rank regression was able to identify candidate genetic variants with novel pleiotropic effects compared to existing methods.

Multiwave sampling for two-phase designs

Thomas Lumley
University of Auckland
E-mail: t.lumley@auckland.ac.nz

Abstract: We consider the problem of estimation in two-phase samples, such as analysing EHR data using a validation subsample. Because of concerns about the impact of even nearly-undetectable model misspecification it is of interest to consider weighted estimation using raked weights (rather than a semiparametric-efficient estimator relying on the assumed model). The optimal sampling design depends on the unknown parameters in the model we want to estimate, and also on on the unknown relationship between the phase 1 and phase 2 data. Following McIsaac and co-workers, we consider multi-wave sampling of the validation sample, where a pilot sample is taken to estimate the parameters and allow the design to be optimised. We consider more complex phase-1 information than previous literature, and also the use of prior distributions.

S133: Statistical advances in accelerating global health and drug development in special population

Application of Model Informed Pediatric Extrapolation in Drug Development

Christine Xu
Sanofi
E-mail: Christine.Xu@sanofi.com

Abstract: Under the regulatory guidance, the pediatric extrapolation concept has been built on evidence of similarities or differences in the disease and the clinical response across the source (e.g. adult) and target populations (e.g. children). Drug development in pediatric patients is challenging. There are many gaps in knowledge need to be investigated under the extrapolation concept and require more innovative approach. The quantitative methods help to establish the relationship between dose, drug exposure, pharmacodynamic effects and clinical responses as well as the understanding of disease. In this session, speaker will demonstrate impacts of model informed extrapolation approaches with a few examples: 1) population pharmacokinetic simulations to support dose selection and optimize the experimental study designs in pediatric population; 2) physiologically-based pharmacokinetic model to extrapolate to a special population and drug-drug interactions; and 3) quantitative systems pharmacology model to integrate knowledge and extrapolate beyond collected data. The applications are varied from more empirical approaches to the integrated computational modeling of biological systems/pharmacologic systems and demonstrate the utility of these quantitative approaches in pediatric drug development.

Janssen's platform trial in evaluating novel compounds in Crohn's Disease

Karen Xia
Johnson and Johnson
E-mail: KXia@its.jnj.com

Abstract: The platform study design concept is proposed as a more efficient means of evaluating novel therapies. The platform study design enables multiple drugs to be evaluated in a clinical study in either a simultaneous or sequential manner. With a solid portfolio in Crohn's Disease, Janssen has designed a phase 2 platform trial to test multiple novel compounds.

We will introduce this case study in two parts. Part 1, the background why Crohn's Disease is selected for Janssen's endeavor on a platform trial, basic platform concept, its major benefits, the protocol structure, and our interactions with regulatory agencies and ethics committees. Part 2, some major differences and challenges associated with our platform study in Crohn's Disease, especially those statistics related, will be shared. These differences include the randomization structure, randomization ratio considerations, Data Monitoring Committee (DMC) set-up, homogeneity evaluations across compounds, data usage in interim analyses and data base locks, among other topics. Some of these differences/challenges are general in platform trials, some of them are specific to Crohn's Disease and Janssen's study design.

Real World Data Informed Clinical Development via Modeling

and Simulation

Zhaoling Meng
Gates MRI
E-mail: zhaoling.meng@gmail.com

Abstract: Historical clinical trials, electronic health records (EHR) and patient population survey provide rich information for the clinical development and enable the project planning going beyond simple assumption-based sample size justification. Modeling and clinical trial simulations (CTS) are frequently utilized to integrate these information with project knowledge and assumptions to provide quantitative assessments and scenario testing in the study and sometime project development design, increase the study/project probability of success (POS) and maximize resource input and outcome benefit balance. This talk presents case studies to illustrate the utilization, especially in complex decision-making situations. First example presented the use of Claims data in supplementing available clinical knowledge and designing a Cardiovascular (CV) safety outcome study. 2nd example illustrated the use of historical study results and targeted disease population survey to run various Phase 3 scenarios and balance the predicted benefit and cost considering a large Phase 2 result uncertainty in a vaccine study setting.

S134: Recent Advances in Machine Learning

LOW-TUBAL-RANK TENSOR RECOVERY FROM ONE-BIT MEASUREMENT

Jianjun Wang
Southwest University
E-mail: wjj@swu.edu.cn

Abstract: Simultaneous control on true positive rate and false positive rate is of significant importance in the performance evaluation of diagnostic tests. Most of the established literature utilizes partial area under receiver operating characteristic (ROC) curve with restrictions only on false positive rate (FPR), called FPR pAUC, as a performance measure. However, its indirect control on true positive rate (TPR) is conceptually and practically misleading. In this paper, a novel and intuitive performance measure, named as two-way pAUC, is proposed, which directly quantifies partial area under ROC curve with explicit restrictions on both TPR and FPR. To estimate two-way pAUC, we devise a nonparametric estimator. Based on the estimator, a bootstrap-assisted testing method for two-way pAUC comparison is established. Moreover, to evaluate possible covariate effects on two-way pAUC, a regression analysis framework is constructed. Asymptotic normalities of the methods are provided. Advantages of the proposed methods are illustrated by simulation and Wisconsin Breast Cancer Data. We encode the methods as a publicly available R package tpAUC.

The power of depth for deep nets in learning theory

Shao-Bo Lin
Xi'an Jiaotong University
E-mail: sblin1983@gmail.com

Abstract: The objective of this talk is to show the power of depth for deep nets in learning theory. In particular, we find that the depth plays an important role for neural networks in providing localized approximation, manifold learning, realizing rotation invariance priors and embodying sparsity in the frequency domain and in the spatial domain . We establish almost optimal learning rates for learning the related functions in a standard learning theory framework.

Estimation of the Number of Endmembers via Thresholding Ridge Ratio Criterion

Xuehu Zhu
Xi'an Jiaotong University
E-mail: zhuxuehu@xjtu.edu.cn

Abstract: Endmember is defined as the spectral signature of pure material present in hyperspectral imagery. Estimation of the number of endmembers (NOE) present in a scene is an important preprocessing step and plays a crucial role in hyperspectral image processing, since over- or under-estimation of the NOE will lead to heavily incorrect results. In this paper, we develop a thresholding ridge ratio (TRR) criterion based on eigendecomposition for NOE determination. Different from the widely used eigenvalue difference analysis methods, the TRR seeks an adaptive thresholding operation to the ridge ratio of eigenvalue differences. And ridge ratio combined with adaptive thresholding can theoretically guarantee a consistent estimate even when there are several local minima. Based on the TRR criterion, an algorithm is introduced to perform the estimation of NOE. Experimental results on both simulated and real hyperspectral data sets have demonstrated that the proposed TRRbased algorithm has comparable and even better performances to several benchmark algorithms in estimation accuracy of the NOE.

A Fast and Accurate Frequent Directions Algorithm for Low Rank Approximation via Block Krylov Iteration

Yao Wang
Xi'an Jiaotong University
E-mail: yao.s.wang@gmail.com

Abstract: It is known that frequent directions (FD) is a popular matrix sketching method for low rank approximation. However, FD and its randomized variants usually meet high computational cost or computational instability in dealing with large-scale datasets, which limits their use in practice. To remedy such issues, this work aims at improving the efficiency and effectiveness of FD. Specifically, by utilizing the power of Block Krylov Iteration and count sketch techniques, we propose a fast and accurate FD algorithm dubbed as BKICS-FD. We analyze the theoretical properties of the proposed BKICS-FD and then carry out extensive numerical experiments to illustrate its superiority over several popular FD algorithms, both in terms of computational speed and accuracy.

S135: Random Matrix Theory and its Applications to Statistics

Eigenvector distribution of deformed random matrices

Xiucai Ding
DUKE UNIVERSITY
E-mail: xiucai.ding@duke.edu

Abstract: In this talk, I will present the recent results on the eigenvector distribution of spiked sample covariance matrices.

ON EIGENVALUES OF A HIGH-DIMENSIONAL SPATIAL-SIGN COVARIANCE MATRIX

Weiming Li
Shanghai University of Finance and Economics
E-mail: li.weiming@shufe.edu.cn

Abstract: This paper investigates limiting properties of eigenvalues of multivariate sample spatial-sign covariance matrices when both the number of variables and the sample size grow to innity. The underlying p-variate populations are general enough to include the popular independent

components model and the family of elliptical distributions. A rst result of the paper establishes that the distribution of the eigenvalues converges to a deterministic limit that belongs to the family of generalized Marcenko-Pastur distributions. Furthermore, a new central limit theorem is established for a class of linear spectral statistics. We develop two applications of these results to robust statistics for a high-dimensional shape matrix. First, two statistics are proposed for testing the sphericity. Next, a spectrum-corrected estimator using the sample spatial-sign covariance matrix is proposed. Simulation experiments show that in high dimension, the sample spatial-sign covariance matrix provides a valid and robust tool for mitigating influence of outliers.

On testing high dimensional white noise

Zeng Li
Southern University of Science and Technology
E-mail: liz9@sustech.edu.cn

Abstract: "Testing for white noise is a classical yet important problem in statistics, especially for diagnostic checks in time series modeling and linear regression. For high-dimensional time series in the sense that the dimension p is large in relation to the sample size T, the popular omnibus tests including the multivariate Hosking and Li-McLeod tests are extremely conservative, leading to substantial power loss. To develop more relevant tests for high-dimensional cases, we propose a portmanteau-type test statistic which is the sum of squared singular values of the rst q lagged sample autocovariance matrices. It, therefore, encapsulates all the serial correlations (upto the time lag q) within and across all component series. Using the tools from random matrix theory and assuming both p and T diverge to innity, we derive the asymptotic normality of the test statistic under both the null and a specic VMA(1) alternative hypothesis. As the actual implementation of the test requires the knowledge of three characteristic constants of the population cross-sectional covariance matrix and the value of the fourth moment of the standardized innovations, non trivial estimations are proposed for these parameters and their integration leads to a practically usable test. Extensive simulation conrms the excellent nite-sample performance of the new test with accurate size and satisfactory power for a large range of nite (p; T) combinations, therefore ensuring wide applicability in practice. In particular, the new tests are consistently superior to the traditional Hosking and Li-McLeod tests."

Can we trust PCA on nonstationary data?

Yanrong Yang
Australian National University
E-mail: yanrong.yang@anu.edu.au

Abstract: This paper investigates the asymptotic distribution of the spiked empirical eigenvalues for high dimensional complicated data, which take into account various structures of the population covariance matrix, dependent sample observations and large dimensionality. It provides new insights into three important roles that play in principal component analysis (PCA): the leading population eigenvalues, dependent sample observations and dimensionality. A surprising discovery is that spiked empirical eigenvalues will reflect the dependent sample structure instead of the population covariance under some scenarios, which indicates possibly inaccurate dimension reduction from PCA for high dimensional data. In particular, we show some modern statistical methods fail in estimating the number of spiked population eigenvalues for high dimensional data with factor model structure and dependent sample observations. To make further study, we propose a test statistic to distinguish spiked population covariance structure from dependent sample structure, especially for high dimensional time series with unit root. Our results are successfully applied to OECD health care expenditure data and US mortality data, which illustrate nonstationary strong temporal dependence. We provide justification for popular literature on mortality forecasting, in which PCA is applied on mortality data directly.

S136: Robust Statistics

Confidence intervals for multiple isotonic regression and other monotone models

Qiyang Han
Rutgers University
E-mail: qh85@stat.rutgers.edu

Abstract: "We consider the problem of constructing pointwise confidence intervals in the multiple isotonic regression model. Recently, Han and Zhang cite{han2019limit} obtained a pointwise limit distribution theory for the max-min block estimator cite{fokianos2017integrated} in this model, but inference remains a difficult problem due to the nuisance parameter in the limit distribution that involves multiple unknown partial derivatives of the true regression function.

In this paper, we show that this nuisance parameter can be effectively eliminated by taking advantage of information beyond point estimates in the max-min block estimator, by establishing a pivotal limiting distribution theory. This immediately yields confidence intervals for $f_0(x_0)$ with asymptotically exact confidence level and optimal length. The construction of the confidence intervals can be easily adapted to other common monotone models including, e.g., (i) monotone density estimation, (ii) interval censoring model with current status data and (iii) counting process model with panel count data. Extensive simulation results demonstrate the accuracy of the coverage probability of the proposed confidence intervals, giving strong support to our theory."

Adaptive Minimax Density Estimation for Huber's Contamination Model under L_p losses

Zhao Ren
University of Pittsburgh
E-mail: zren@pitt.edu

Abstract: Today's data pose unprecedented challenges as it may be incomplete, corrupted or exposed to some unknown source of contamination. In this talk, we address the problem of density function f estimation under L_p losses ($1leq p <infty$) for Huber's contamination model in which one observes i.i.d. observations from $(1-epsilon)f+epsilon g$ and g represents the unknown contamination distribution. We investigate the effects of contamination proportion $epsilon$ among other key quantities on the corresponding minimax rates of convergence for both structured and unstructured contamination classes: for structured contamination, $epsilon$ always appears linearly in the optimal rates while for unstructured contamination, the leading term of the optimal rate involving $epsilon$ also relies on the smoothness of target density class and the specific loss function.

We further carefully study the corresponding adaptation theory in contamination models. Two different Goldenshluger-Lepski-type methods are proposed to select bandwidth and achieve L_p risk oracle inequalities for structured and unstructured contaminations respectively. It is shown that the proposed procedures lead to minimax rate-adaptivity over a scale of the

anisotropic Nikol'skii classes for most scenarios except that adaptation to both contamination proportion $epsilon$ and smoothness of density class for unstructured contamination is shown to be impossible. Our technical analysis in adaptive procedures relies on some uniform bounds under the L_p norm of empirical processes developed by Goldenshluger and Lepski.

Consistency of a range of penalised cost approaches for detecting multiple changepoints

Chao Zheng
Lancaster University
E-mail: c.zheng5@lancaster.ac.uk

Abstract: A common approach to detect multiple changepoints is to minimise a measure of data fit plus a penalty that is linear in the number of changepoints. In this talk, we show that the general finite sample behaviour of such a method can be related to its behaviour when analysing data with either none or one changepoint. This results in simpler conditions for verifying whether the method will consistently estimate the number and locations of the changepoints. We apply and demonstrate the usefulness of this result for a range of changepoint problems. Our new results include a weaker condition on the choice of penalty required to have consistency in a change-in-slope model; and the first results for the accuracy of recently-proposed methods for detecting spikes.

On Perfect Classification and Clustering for Gaussian Processes

Subhajit Dutta
IIT KANPUR
E-mail: tijahbus@gmail.com

Abstract: According to the Hajek-Feldman property, two Gaussian distributions are either equivalent or mutually singular in the infinite-dimensional case. Motivated by singularity of a class of Gaussian measures, we first state a result based on the classic Mahalanobis distance and give an outline of the proof. Using this basic result, a joint transformation is proposed and its theoretical properties are investigated. In a classification problem, this transformation induces complete separation among the competing classes and a simple component-wise classifier leads to 'perfect classification' in such scenarios. In the second part of this talk, we shall discuss the problem of identifying groups in a mixture of Gaussian processes (clustering) by using a new transformation involving Mahalanobis distances. It is curious to note that the proposed method is useless in homoscedastic cases, however, it yields 'perfect clustering' for groups having differences in their covariance operators.

(a joint work with Prof. Juan A. Cuesta-Albertos)

S137: Causal inferences in survival and mediation analyses

Instrumental variable estimation of a Cox marginal structural model with time-varying endogenous treatments

Yifan Cui
University of Pennsylvania
E-mail: cuiy@wharton.upenn.edu

Abstract: Robins (1998) introduced marginal structural models (MSMs), a general class of counterfactual models for the joint effects of time-varying treatment regimes in complex longitudinal studies subject to time-varying confounding. He established the identification of MSM parameters under a sequential randomization assumption (SRA), which rules out unmeasured confounding of treatment assignment over time. The Cox marginal structural model, in particular, is one of the most popular MSMs for evaluating the causal effect of a binary exposure with a censored failure time outcome. In this paper, we consider sufficient conditions for identification of Cox MSM parameters with the aid of a time-varying instrumental variable, when sequential randomization fails to hold due to unmeasured confounding. Our identification conditions essentially require that no interactions between unmeasured confoundings and the instrumental variable in its additive effects on the treatment, the longitudinal generalization of the identifying condition of Wang et al. (2018). Our approach is illustrated via extensive simulation studies and real data example.

Debiased Inverse-Variance Weighted Estimator in Two-Sample Summary-Data Mendelian Randomization

Ting Ye
University of Pennsylvania
E-mail: tingye@wharton.upenn.edu

Abstract: Recently, Mendelian randomization has become a popular approach to study the effect of a modifiable biomarker or exposure on an outcome of interest using genetic variants from pre-existing genome-wide association studies as instruments. A challenge of using genetic variants as instruments is that each individual genetic variant usually explains a relatively small proportion of variance in the exposure and there are many such instruments, a setting known as many weak instruments. Unfortunately, some popular estimators in Mendelian Randomization are developed under the strong instruments setting and only empirical studies have shown that they are biased under the many weak instruments setting. In this paper, we study the theoretical properties of the two most popular estimators in Mendelian Randomization, the inverse-variance weighted (IVW) estimator and pre-screened IVW estimator using strong instruments selected from a selection dataset. We provide a full characterization of these estimators with many weak instruments by using a measure of average instrument strength. Based on our theoretical investigations, we propose a debiased IVW estimator, a simple modification of the IVW estimator, that is robust to many weak instruments and requires no pre-screening. When a selection dataset is available, we propose two principled ways to determine the p-value cutoff for pre-screening to improve efficiency of the debiased IVW estimator. An extension of debiased IVW estimator to the balanced horizontal pleiotropy is also discussed. We conclude by demonstrating our results in simulated and real datasets.

Causal mediation analysis with multiple causally non-ordered mediators

Masataka Taguri
Yokohama City University
E-mail: taguri@yokohama-cu.ac.jp

Abstract: In many health studies, researchers are often interested in estimating the treatment effects on the outcome around and through an intermediate variable, where the two effects are called direct and indirect effects respectively and add to the total treatment effect. Such causal mediation analyses aim to understand the mechanisms that explain the treatment effect. Although multiple mediators are often involved in real studies, most of the literature considered mediation analyses with one mediator at a time. In this presentation, we consider mediation analyses when there are causally non-ordered multiple mediators. Even if the

mediators do not affect each other, the sum of two indirect effects through the two mediators considered separately may diverge from the joint natural indirect effect of them when there are additive interactions between the effects of the two mediators on the outcome. Therefore, we derive an equation for the joint natural indirect effect based on the individual mediation effects and their interactive effect, which helps us understand how the mediation effect works through the two mediators. We also discuss an extension for three mediators. The proposed method is illustrated using data from a randomized trial on the prevention of dental caries.

The randomization distribution of the logrank statistic

Xinran Li
University of Illinois at Urbana-Champaign
E-mail: xinranli@illinois.edu

Abstract: The logrank test is one of the most popular approaches for comparing time-to-event outcomes in the presence of censoring. Most theoretical justifications given for it have required a hypothetical superpopulation, in the sense that the event times for all units are independent and identically distributed (i.i.d.). We invoke the potential outcome framework to define the causal effect of certain treatment on a time-to-event outcome, and conduct finite population inference that relies crucially on the physical randomization of the treatment. On the one hand, finite population inference focuses particularly on the finite experimental units by viewing their potential outcomes as fixed constants; this is equivalent to conducting inference conditioning on all the potential outcomes. On the other hand, the test justified by finite population inference can be valid without any distributional assumptions (such as i.i.d.) on the potential outcomes. In this paper, we study finite population inference for the logrank test, and specifically we investigate the randomization distribution of the logrank statistic. We show that, under a Bernoulli randomized experiment with non-informative i.i.d. censoring within each treatment arm, the logrank test is asymptotically valid for testing Fisher's null hypothesis of no treatment effect on any unit. The asymptotic validity of the logrank test does not require any distributional assumptions on the potential event times; for example, the potential event times can have arbitrary dependence and heterogeneity across the units. The developed theory for the logrank rank test from finite population inference supplements its classical theory from usual superpopulation inference, and thus helps provide a broader justification for the logrank test. We also extend the theory to the stratified logrank test, which is useful for randomized blocked designs and when censoring mechanisms vary across strata.

S139: Time Series Analysis

Modeling Financial Time Series with Soft Information

Shih-Feng Huang
National University of Kaohsiung
E-mail: huangsf@nuk.edu.tw

Abstract: A hysteretic autoregressive model with GARCH effects and soft information, denoted by SHAR-GARCH, is proposed to model financial time series. The soft information contained in the daily news is extracted by the techniques of support vector machine and principal component analysis. A Markov Chain Monte Carlo algorithm is proposed for estimating model parameters. A corresponding risk-neutral SHAR-GARCH model is derived by Esscher transform for option pricing. The returns and options of the S&P500 index and the daily news posted on the website of Reuters are used for our empirical study. The numerical results indicate that the proposed model has satisfactory performances in depicting the dynamics of financial time series and in pricing deep-in-the-money options.

Time Series Analysis with Unsupervised Learning

Meihui Guo
Dept. of Applied Mathematics, National Sun Yat-sen University
E-mail: guomh@math.nsysu.edu.tw

Abstract: We consider the prediction problem for time series with unknown clusters. Unsupervised learning methods, such as hierarchical and K-means clustering techniques are applied to pre-cluster the time series trend. Non-parametric approaches are adopted to estimate the trends of the clusters. We use conventional time series models and long short-terms memory network (LSTM) models to fit the original and de-trended time series data and compare their prediction performance. In the empirical study, we analyze daily/intra-daily mass transit vehicle capacity time series data of Kaohsiung city. The results show that the conventional time series models have better prediction performance than the LSTM model for stationary case, yet the LSTM models perform better for non-stationary case including change point.

Symbolic Interval-Valued Data Analysis for Time Series Based on Auto-interval-regressive Models

Liang-Ching Lin
National Cheng Kung University
E-mail: lclin@mail.ncku.edu.tw

Abstract: This study considers interval-valued time series data. To characterize interval time series data, we propose an auto-interval-regressive (AIR) model using the order statistics from normal distributions. Furthermore, to better capture heteroscedasticity in volatility, we designate a heteroscedastic volatility auto-interval-regressive (HVAIR) model. We derive the likelihood functions of the AIR and HVAIR models to obtain the maximum likelihood estimator. Monte Carlo simulations are then conducted to evaluate our methods of estimation and conrm their validity. A real data example from the S&P 500 Index is used to demonstrate our method.

Clt for largest eigenvalues in high-dimensional nonstationary time series and its applications

Bo Zhang
University of Science and Technology of China
E-mail: zhangbo890301@outlook.com

Abstract: This paper considers a p-dimensional non-stationary time series model. We investigate the asymptotic behavior of the first k largest eigenvalues of the sample covariance matrices of the time series model. Then we propose an estimator of autoregressive coefficient and use it to test the near unit root. The convergence rate of the estimator is better than existed methods. Simulations are also conducted to demonstrate the performances of the estimator and the statistic.

S140: High dimensional change point detection

A Composite Likelihood-based Approach for Change-point Detection in Spatio-temporal Process

Chun Yip Yau
Chinese University of Hong Kong
E-mail: cyyau@sta.cuhk.edu.hk

Abstract: This paper develops a unified, accurate and computationally efficient method for change-point inference in non-stationary

spatio-temporal processes. By modeling a non-stationary spatio-temporal process as a piecewise stationary spatio-temporal process, we consider simultaneous estimation of the number and locations of change-points, and model parameters in each segment.

A composite likelihood-based criterion is developed for change-point and parameters estimation.

Asymptotic theories including consistency and distribution of the estimators are derived under mild conditions.

In contrast to classical results in fixed dimensional time series that the asymptotic error of change-point estimator is $O_{p}(1)$, exact recovery of true change-points is guaranteed in the spatio-temporal setting.

More surprisingly, the consistency of change-point estimation can be achieved without any penalty term in the criterion function.

A computational efficient pruned dynamic programming algorithm is developed for the challenging criterion optimization problem.

Simulation studies and an application to U.S. precipitation data are provided to demonstrate the effectiveness and practicality of the proposed method.

Distributed linear regression in high dimensions

Yue Sheng
University of Pennsylvania
E-mail: yuesheng@sas.upenn.edu

Abstract: Distributed statistical learning problems arise commonly when dealing with large datasets. In this setup, datasets are partitioned over machines, which compute locally and communicate short messages. Communication is often the bottleneck. In this paper, we study one-step and iterative

weighted parameter averaging in statistical linear models under data parallelism. We do linear regression on each machine, send the results to a central server, and take a weighted average of the parameters. Optionally, we iterate, sending back the weighted average and doing local ridge regressions centered at it. How does this work compare to doing linear regression on the

full data? Here we study the performance loss in estimation and test error, and confidence interval length in high dimensions, where the number of parameters is comparable to the training data size.

We find the performance loss in one-step weighted averaging, and also give results for iterative averaging. We also find that different problems are affected differently by the distributed framework. Estimation error and confidence interval length increases a lot, while the prediction error increases much less.

High dimensional clustering

Guangming Pan
Nanyang Technological University
E-mail: stapgm@gmail.com

Abstract: This talk is about high dimensional clustering. We propose a two step method for mixture data by random matrix theory when the sample size and the dimension of the data are comparable to each other.

Selection of the number of change-points via error rate control

Changliang Zou
Nankai University
E-mail: nk.chlzou@gmail.com

Abstract: In multiple change-point analysis, one of the main difficulties is to determine the number of change-points. Various consistent selection methods, including the use of Schwarz information criterion or cross-validation have been proposed to balance the model fitting and complexity. However, there is a lack of systematic approach to providing ``significance" information in determining the number of changes. I will introduce a data-adaptive selection procedure via error rate control, which is applicable to most kinds of popular change-point algorithms. The key idea is to apply the order-preserved sample-splitting strategy to construct a series of statistics with marginal symmetry property and then to utilize the symmetry for constructing a data-driven threshold. The false discovery rate (FDR) control is detailedly investigated and some other error rates are also discussed. We show that the proposed method is able to, at least asymptotically, control the FDR under certain conditions and still retain all of the true change-points. Some important examples are presented to illustrate the merits of our procedure. Numerical experiments indicate that the proposed methodology works well for many existing change-detection methods and is able to yield accurate FDR control in finite samples.

S141: Recent Progresses on Dimension Reduction & High Dimensional Data Analysis

Sparse SIR via Lasso

Qian lin
Tsinghua University
E-mail: qianlin@mail.tsinghua.edu.cn

Abstract: For multiple index models, it has recently been shown that the sliced inverse regression (SIR) is consistent for estimating the sufficient dimension reduction (SDR) space if and only if ρ=limpn=0, where p is the dimension and n is the sample size. Thus, when p is of the same or a higher order of n, additional assumptions such as sparsity must be imposed in order to ensure consistency for SIR. By constructing artificial response variables made up from top eigenvectors of the estimated conditional covariance matrix, we introduce a simple Lasso regression method to obtain an estimate of the SDR space. The resulting algorithm, Lasso-SIR, is shown to be consistent and achieve the optimal convergence rate under certain sparsity conditions when p is of order o(n2λ2), where λ is the generalized signal-to-noise ratio. We also demonstrate the superior performance of Lasso-SIR compared with existing approaches via extensive numerical studies and several real data examples.

Bayesian Sufficient Dimension Reduction via Modeling Joint Distributions

Yingkai Jiang
Tsinghua University
E-mail:ykjiang15@gmail.com

Abstract: This work develops a novel method to estimate the central subspace in the sufficient dimension reduction problem. By modeling the joint distribution of the projected predictive variables and the response variable, we can assess the likelihood of a specific projection subspace and the corresponding joint distribution, thus estimate both simultaneously. The main difficulty of this approach is that the parameter space is extremely large. Therefore, the Markov chain converges slowly. We develop an efficient sampling method for this model and enable the computation time

comparable with other iterative methods. This joint modeling approach is capable of detecting complicated relationships between predictors and the response. And we can do posterior inference rather than just point estimations in the Bayesian framework. These advantages are confirmed in simulation studies.

Multiple influential point detection in high dimensional regression spaces

Junlong Zhao
School of Statistics, Beijing Normal University
E-mail: zhaojunlong928@126.com

Abstract: Influence diagnosis is an integrated component of data analysis but has been severely underinvestigated in a high dimensional regression setting. One of the key challenges, even in a fixed dimensional setting, is how to deal with multiple influential points that give rise to masking and swamping effects. The paper proposes a novel group deletion procedure referred to as multiple influential point detection by studying two extreme statistics based on a
marginal-correlation-based influence measure. Named the min- and max-statistics, they have complementary properties in that the max-statistic is effective for overcoming the masking effect whereas the min-statistic is useful for overcoming the swamping effect. Combining their strengths, we further propose an efficient algorithm that can detect influential points with a prespecified false discovery rate. The influential point detection procedure proposed is simple to implement and efficient to run and enjoys attractive theoretical properties. Its effectiveness is verified empirically via extensive simulation study and data analysis. An R package implementing the procedure is freely available.

S142: Promoting Statistical Consulting and Collaboration in China

Effective Communication for Successful Collaboration

Xiaoyue Niu
the Pennsylvania State University
E-mail: xiaoyue@psu.edu

Abstract: Statistics is an interdisciplinary field. Successful collaboration can solve practically important questions and inspire new statistical methodology. Communication skills are critical in all collaboration. In this talk, I will introduce some basic communication skills and discuss how they can lead to a successful collaboration.

Statistical Consulting in the Era of Data Science

Lillian Lin
Retired Statistical Consultant
K-mail: ls.lin.mt@gmail.com

Abstract: Statistics emerged from the examination of issues confronted by genetics, betting, and agriculture and is now an essential tool for nearly every scientific field. Yet, there has never been a focused interdisciplinary examination of how to construct a framework for statistical collaboration that satisfies the intellectual values of experts from fields of application as well as the statistics profession. Rather, these experts are satisfied to learn statistical recipes (e.g., how to use software for analyses typically deemed acceptable in their peer reviewed journals) rather than seeking out a statistician who can take responsibility for a study design and data analyses tailored to their specific investigation. Statisticians, in turn, have retreated to inventing ever more specialized methods (an approach supported by our peer reviewed journals) which results in ceding control of the definition of appropriate statistical collaboration to non-statisticians. In recent years, this problem has been compounded by the emergence of the field of data science. Statisticians are often pushed to perform technical tasks that stray far beyond our expertise and fail to receive proper credit for intellectual contributions. Drawing on decades of experience serving as a statistician collaborating with scientists and, more recently, in creating an academic statistical consulting service, I will propose and examine issues that every statistical consulting service, whether an individual consultant or an organization, must address in order to preserve intellectual identity for statisticians and our scientific colleagues and promote effective quantitative research.

Strategies for promoting the engagement of students in statistical consulting

Ximing Xu
Nankai University
E-mail: nk_xu@hotmail.com

Abstract: Students can not only practice what they have learned by participating in statistical consulting , but also learn new knowlege and essential skills for being a qualified statistician. In this talk some strategies to increase the interest and engagement of students based on the practice at Nankai Unversity will be introduced, with projects on medical data used as illustrative examples.

Statistical Consulting Practice at Tsinghua University

Mengzhao Gao
Tsinghua University
E-mail: mog5192@mail.tsinghua.edu.cn

Abstract: Statistical consulting practice has been established and proven to be an effective approach in statistical education at many universities in U.S. and Europe. In China, Center for Statistical Science of Tsinghua University has been exploring how to provide data and statistical analysis and consulting services for teachers and students of different disciplines inside the school, as well as for clients from industries and governments. Through this talk, I will introduce Statistical Consulting Center at Tsinghua University from the perspective of operations and management, and share our experience of collaborating with clients through case-studies.

S143: Statistical Advances and Challenges in Bioinformatics

miRACLe: improving the prediction of miRNA-mRNA interactions by a random contact model

Qi Li
Tsinghua University
E-mail: liqi123124@163.com

Abstract: The strength of miRNA-mRNA interactions in a biological system depends on both the sequence characteristics and expression patterns of RNAs. Integrating the two features into a random contact model, we propose miRACLe (miRNA Analysis by a Contact modeL) to achieve effective miRNA target prediction at both individual and population levels. Evaluation by a variety of measures shows that fitting a sequence-based algorithm into the framework of miRACLe can improve its predictive power with a significant margin, and the combination of miRACLe and TargetScan consistently outperforms state-of-the-art methods in prediction accuracy, regulatory potential and biological relevance.

Quantifying the impact of genetically regulated expression on complex traits and diseases

Can Yang
HKUST
E-mail: macyang@ust.hk

Abstract: By leveraging existing GWAS and eQTL resources, transcriptome-wide association studies (TWAS) have achieved many successes in identifying trait-associations of genetically-regulated expression (GREX) levels. TWAS analysis relies on the shared GREX variation across GWAS and the reference eQTL data, which depends on the cellular conditions of the eQTL data. Considering the increasing availability of eQTL data from different conditions and the often unknown trait-relevant cell/tissue-types, we propose a method and tool, IGREX, for precisely quantifying the proportion of phenotypic variation attributed to the GREX component. IGREX takes as input a reference eQTL panel and individual-level or summary-level GWAS data. Using eQTL data of 48 tissue types from the GTEx project as a reference panel, we evaluated the tissue-specific IGREX impact on a wide spectrum of phenotypes. We observed strong GREX effects on immune-related protein biomarkers. By incorporating trans-eQTLs and analyzing genetically-regulated alternative splicing events, we evaluated new potential directions for TWAS analysis.

Statistical Analysis of Somatic Mutations in Cancer Genomes

Wei Sun
Fred Hutch
E-mail: wsun@fredhutch.org

Abstract: Somatic mutations drive the growth of tumor cells and are pivotal biomarkers for many cancer treatments. In contrast to germline mutations, somatic mutations may occur in a subset of tumor cells (intra-tumor heterogeneity) and calling somatic mutations often have non-ignorable false positive rate and/or false negative rate. I will present our recent on association analysis using somatic mutations, while accounting for somatic mutation calling uncertainty, and association analysis for intra-tumor heterogeneity.

Dimension Reduction and Dropout Imputation for Single-Cell RNA Sequencing Data Using Constrained Robust Nonnegative Matrix Factorization

Shuqin Zhang
Fudan University
E-mail: zhangs@fudan.edu.cn

Abstract: Single cell RNA-sequencing (scRNA-seq) technology is a powerful tool to analyze the whole transcriptome at single cell level, and it has been receiving more and more attention in recent years. Dimension reduction and clustering are the basic steps in scRNA-seq data analysis, and they are seriously affected by the dropout phenomenon, which is an important characteristic of scRNA-seq data. The cells with lower sequencing depth will tend to have more dropouts compared with the deeper sequenced cells. Moreover, scRNA-seq data is count-based and thus nonnegative.

In this paper, we propose a model for simultaneously implementing dropout imputation and dimension reduction of scRNA-seq data under the nonnegative matrix factorization (NMF) framework.

The dropouts modeled as a nonnegative sparse matrix are added to the observed data matrix, which is approximated by NMF. A weighted L_1 penalty taking into account the dependence of the dropouts on the sequencing depth in each single cell is imposed to ensure the sparsity pattern. The computational efficient method is developed to solve the formulated optimization problem. Experiments on both synthetic data and real data show that the dimension reduction can give more robust clustering results compared with the existing methods and the dropout imputation helps improve the differential expression analysis.

S144: Bayesian Statistics

Efficient Bernoulli factory MCMC for intractable likelihoods

Dootika Vats
Indian Institute of Technology
E-mail: dootika@iitk.ac.in

Abstract: Accept-reject based Markov chain Monte Carlo (MCMC) algorithms have traditionally been a function of the ratio of the target density at the two contested points. We note that this feature is rendered almost useless in Bayesian MCMC problems with intractable likelihoods. We introduce a new acceptance probability that has the distinguishing feature of not being a function of the ratio of the target density at two points. We show that such a structure allows for the construction of an efficient and stable Bernoulli factory. The resulting Portkey Barker's algorithm is exact and computationally more efficient that the current state-of-the-art.

Estimating densities with nonlinear support using Fisher-Gaussian kernels

Minerva Mukhopadhyay
Indian Institute of Technology Kanpur
E-mail: minervamukherjee@gmail.com

Abstract: Current tools for multivariate density estimation struggle when the density is concentrated near a nonlinear subspace or manifold. Most approaches require choice of a kernel, with the multivariate Gaussian kernel being the most commonly used one. Although heavy-tailed and skewed extensions have been proposed, such kernels cannot capture curvature in the support of the data. Hence, a very large number of kernels may be needed to provide an adequate fit to many datasets. This leads to poor performance unless the sample size is very large relative to the dimension of the data, even in toy problems. With this motivation, we propose a novel generalization of the Gaussian distribution, which includes an additional curvature parameter. We refer to the proposed class as Fisher-Gaussian (FG) kernels, since they arise by sampling data from a von Mises-Fisher density on the sphere and adding Gaussian noise. The FG density has an analytic form, and is amenable to straightforward implementation within Bayesian mixture models using Markov chain Monte Carlo. We provide asymptotic theory on posterior concentration, and illustrate gains relative to current methods on simulated and real data applications.

Bayesian Analysis with Gaussian Random Functional Dynamic Spatio-Temporal Model

Suman Guha
Department of Statistics, Presidency University
E-mail: suman.stat@presiuniv.ac.in

Abstract: Discrete-time spatial time series data arise routinely in meteorological and environmental studies. Inference and prediction associated with them are mostly carried out using any of the several variants of the linear state space model that are collectively called linear dynamic spatio-temporal models (LDSTMs). However, real world environmental processes are highly complex and are seldom representable by models with such simple linear structure. Hence, nonlinear dynamic spatio-temporal models (NLDSTMs) based on the idea of nonlinear observational and evolutionary equations have been proposed as an alternative. However, in

that case, the caveat lies in selecting the specific form of nonlinearity from a large class of potentially appropriate nonlinear functions. Moreover, modeling by NLDSTMs requires precise knowledge about the dynamics underlying the data. In this article, we address this problem by introducing the Gaussian random functional dynamic spatio-temporal model (GRFDSTM). Unlike the LDSTMs or NLDSTMs, in GRFDSTM both the functions governing the observational and evolutionary equations are composed of Gaussian random functions. We exhibit many interesting theoretical properties of the GRFDSTM and demonstrate how model fitting and prediction can be carried out coherently in a Bayesian framework. We also conduct an extensive simulation study and apply our model to real datasets. The results are highly encouraging.

S145：Special Invited Papers of Statistics and Its Inference

Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data

Ling Zhou
Southwestern University of Finance and Economics
E-mail: zhouling@swufe.edu.cn

Abstract: The linear mixed-effects model (LMM) is widely used in the analysis of clustered or longitudinal data. This paper aims to address analytic challenges arising from estimation and selection in the application of the LMM to high-dimensional longitudinal data. We develop a doubly regularized approach in the LMM to simultaneously select fixed and random effects. On the theoretical front, we establish large sample properties for the proposed method under the high-dimensional setting, allowing both numbers of fixed effects and random effects to be much larger than the sample size. We present new regularity conditions for the diverging rates, under which the proposed method achieves both estimation and selection consistency. In addition, we propose a new algorithm that solves the related optimization problem effectively so that its computational cost is comparable with that of the Newton-Raphson algorithm for maximum likelihood estimator in the LMM. Through simulation studies we assess performances of the proposed regularized LMM in both aspects of variable selection and estimation. We also illustrate the proposed method by two data analysis examples.

Authors: Yun Li, Sijian Wang, Peter X.-K. Song, Naisyin Wang, Ling Zhou, and Ji Zhu

Optimal Treatment Assignment of Multiple Treatments with Analysis of Variance Decomposition

Zhilan Lou
Zhejiang University of Finance and Economics
E-mail: louzhilan@126.com

Abstract: Personalized medicine to identify individualized treatment assignment rules has received increasing interest. When there are more than two treatments, the outcome weighted learning framework builds an optimal assignment rule via the skill of reproducing kernel Hilbert space. One main challenge is that the interpretation of covariates is difficult since the solution is a black-box classifier. Consequently, we establish a structured optimal treatment assignment rule with the functional analysis of variance decomposition. The method promotes the sparsity of the final solution by using structured kernel function and an l_1 penalty term. Meanwhile, we propose an easy-handling iterative procedure to overcome the calculation problem. Convergence of the risk function for resulting estimator is shown in the paper. The finite sample performance of the proposed method is demonstrated by simulation studies and a real data analysis.

Bayesian modeling and uncertainty quantification for descriptive social networks

Sudipto Banerjee
University of California Los Angeles
E-mail: sudipto@ucla.edu

Abstract: This talk presents an easily implementable Bayesian approach to model and quantify uncertainty in small descriptive social networks. While statistical methods for analyzing networks have seen burgeoning activity over the last decade or so, ranging from social sciences to genetics, such methods usually involve sophisticated stochastic models whose estimation requires substantial structure and information in the networks. At the other end of the analytic spectrum, there are purely descriptive methods based upon quantities and axioms in computational graph theory. In social networks, popular descriptive measures include, but are not limited to, the so called Krackhardt's axioms. Another approach, recently gaining attention, is the use of PageRank algorithms. While these descriptive approaches provide insight into networks with limited information, including small networks, there is, as yet, little research detailing a statistical approach for small networks. This talk presents an interface of Bayesian statistical inference and social network analysis by offering practicing social scientists a relatively straightforward Bayesian approach to account for uncertainty while conducting descriptive social network analysis. The emphasis is on computational feasibility and easy implementation using existing R packages that are available from the Comprehensive R Archive Network (https://cran.r-project.org/). We analyze a network comprising 18 websites from the US and UK to discern transnational identities, previously analyzed using descriptive graph theory with no uncertainty quantification, using fully Bayesian model-based inference.

Analyzing Beijing Point of Interest Data Using Group Linked Cox Process

Yu Chen
Peking University
E-mail: chenyu414@gmail.com

Abstract: We develop in this article a group linked Cox process model for analyzing large-scale data on the point of interest (POI). Our methodology is motivated by a real POI dataset, which contains more than 22 thousand POIs in Beijing urban area. These POIs have been divided into many small categories (e.g., restaurants, movie theaters, hospitals, universities and subway stations) by the digital map maker (e.g., Baidu Map). Empirical analysis provides substantial evidence that POIs across different categories could be highly correlated so that those small categories can be further grouped. To this end, we develop here a group linked Cox process model. Specifically, within each group, we model POI locations by a standard Cox process so that the POI clustering effect can be well described. Furthermore, the idea of bivariate linked Cox process is borrowed and further extended to its multivariate counterpart. Consequently, a more sig- nificant number of POI categories can be accommodated within each group. To estimate the model, a minimum contrast type method is developed, and an au- tomatically grouping method is provided. Simulation studies are conducted to validate the proposed methodology. At last, we apply our method to the

afore- mentioned real dataset, and a total of 4 groups are uncovered. This leads to the discovery of some urban-planning-related features.

S146: Adanced Statistical Methods for Microbiome Sequencing Data with Applications to Complex Human Diseases

Evaluation of Statistical Methods for Differential Expression Analysis in Microbiome Metatranscriptomics Data

Di Wu
University of North Carolina at Chapel Hill
E-mail: dwu@unc.edu

Abstract: Differential abundance (DA) analysis in metagenomics (from whole genome shortgun DNAseq of bacteria) and differential expression (DE) in metatranscriptomics (from RNAseq of bacteria) are critical means for understanding differences between micorbiome sample groups (e.,g disease vs normal). However, modeling microbiome data is challenging because of their sparsity (i.e., zero inflation), over-dispersion, high dimensionality, and their inherent hierarchical compositional structure.

Although various statistical testing methods have been used in DA analysis at the taxa level of metagenomics data, how their performance in DE analysis of metatranscriptomics remains unclear. Therefore a comprehensive evaluation of these methods is required regarding metatranscriptomics data at the bacteria gene level may be more zero-inflated and dispersed. We here symmetrically evaluated five methods including logistic beta of compositional data and a few rank-based methods, using simulations based on the real-data inspired zero inflated negative binomial distribution, with different scenarios in terms of mean shift, dispersion, zero percentage and batch effects. Both Type 1 error rate and power are considered. We will use the evaluation results to guide the usage of methods for metatranscriptomics DE analysis.

Microbial group association test based on the higher criticism

Ni Zhao
Johns Hopkins University
E-mail: nzhao10@jhu.edu

Abstract: In human microbiome studies, it is essential to evaluate the association between microbial group (e.g., community or clade) composition and a host phenotype of interest. In response, a number of microbial group association tests have been proposed, which take into account the unique features of the microbiome data (e.g., high-dimensionality, compositionality, phylogenetic relationship). These tests generally fall in the class of aggregation tests which amplify the overall group association by combining all the underlying microbial association signals; therefore, they are powerful when many microbial species are associated (i.e., low sparsity). However, in practice, the microbial association signals can be highly sparse, and this is the situation where we have a difficulty to discover the microbial group association. Hence, here we introduce a powerful microbial group association test for sparse microbial association signals, namely, microbiome higher criticism analysis (MiHC). MiHC is a data-driven optimal test taken in a search space spanned by tailoring the higher criticism test to incorporate phylogenetic information and/or modulate sparsity levels. Our simulations show that MiHC maintains a high power at different phylogenetic relevance and sparsity levels with correct type I error controls. We also demonstrate the use of MiHC with tree real data applications.

Multi-SNP mediation intersection-union test

Xiaojing Zheng
UNC-CH
E-mail: xiaojinz@email.unc.edu

Abstract: Tens of thousands of reproducibly identified GWAS (Genome-Wide Association Studies) variants, with the vast majority falling in non-coding regions resulting in no eventual protein products, call urgently for mechanistic interpretations. Although numerous methods exist, there are few, if any methods, for simultaneously testing the mediation effects of multiple correlated SNPs via some mediator (e.g. the expression of a gene or a cytokine in the neighborhood) on phenotypic outcome (including microbiome). We propose multi-SNP mediation intersection-union test (SMUT) to fill in this methodological gap. Our extensive simulations demonstrate the validity of SMUT as well as substantial, up to 92%, power gains over alternative methods. In addition, SMUT confirmed known mediators in a real dataset of Finns for plasma adiponectin level, which were missed by many alternative methods. We believe SMUT will become a useful tool to generate mechanistic hypotheses underlying GWAS variants, facilitating functional follow-up.

S147: Statistical Methods and Algorithms for High-dimensional Biomedical Data

A Fast and powerful eQTL weighted method to detect genes associated with complex trait using GWAS summary data

Xuexia Wang
UNIVERSITY OF NORTH TEXAS
E-mail: Xuexia.Wang@unt.edu

Abstract: Although genome-wide association studies (GWASs) have identified many genetic variants underlying complex traits, a large fraction of heritability still remains unexplained. Integrative analysis that incorporates additional information such as expression quantitative trait locus (eQTL) data into sequencing studies (denoted as transcriptome-wide association study TWAS) can aid the discovery of trait associated genetic variants. However, general TWAS methods only incorporate one eQTL-derived weight (e.g. cis-effect), and thus can suffer substantial loss of power when the single estimated cis-effect is not predictive for the effect size of a genetic variant or there are estimation errors in the estimated cis-effect, or even if the data is not consistent with the model assumption. In this study, we propose an omnibus test which utilizes a Cauchy association test to integrate association evidence demonstrated by three different traditional tests (Burden test, Quadratic test and Adaptive test) using GWAS summary data with multiple eQTL-derived weights. The p value of the proposed test can be calculated analytically, and thus it is fast and efficient. We applied our proposed test to two schizophrenia (SCZ) GWAS summary datasets and two lipids trait (HDL) GWAS summary datasets. Compared to the three traditional tests, our proposed omnibus test can identify more trait-associated genes.

Comparisons and Validation of the Three Pulmonary Nodule Malignancy Risk Models (Brock, Radiomics, Deep Learning): A Secondary Analysis of Data from the National Lung Screening Trial

Fenghai Duan
Brown University
E-mail: fduan@stat.brown.edu

Abstract: Lung cancer is the leading cause of cancer death worldwide each year. Non-invasive medical image technologies are becoming routine in

screening high-risk populations for lung cancer patients. Different from the traditional radiological imaging analysis (e.g., manually interpreted by radiologists), the rapid progress of computational methods and artificial intelligence is leading to the extensive implementation of radiomics and/or deep learning in the medical image analysis. In particular, radiomics is the high-throughput extraction and analysis of the quantitative features from advanced medical images with the assistance from compute science to provide a comprehensive quantification of tumor phenotype for cancer patients, and deep learning generally refers to the application of deep neural networks (e.g., CNNs) to process and analyze medical images. In the past we have developed a panel of radiomic features to classify the benign from malignant lung nodules to reduce the unnecessary diagnostic interventions. In this study, we further analyzed the clinical and imaging data from the National Lung Screening Trial and constructed three risk models for discrimination of lung nodule malignancy: the Brock model built from the semantic features, our radiomic model built from the radiomic features, and the deep learning model built from transfer learning. The performance of the three models was thoroughly assessed and compared. In addition, external validation of using an independent cohort has been conducted to evaluate the accuracy of our models.

Double Deep Learning for Adjusting Complex Confounding Structures In Observational Data

Fei Zou
UNC-CH
E-mail: feizou@email.unc.edu

Abstract: Complex confounding structures are often embedded in observational data, including electronic medical record (EMR) data. A robust yet efficient double deep learning approach is proposed to adjust for the complex confounding structures in comparative effectiveness analysis of EMR data. Specifically, deep neural networks are employed to estimate the conditional expectations of the outcome and the treatment allocation given observed baseline covariates under a semiparametric framework. An improved estimation scheme is further developed to enhance the finite sample performance of the proposed method. Comprehensive numerical studies have shown superior performance of the proposed method, as compared with other existing methods, in terms of reduced bias and mean squared error of the treatment effect estimate.

S148: Statistical Issues in Imaging Data Analysis

Bayesian Spatial Blind Source Separation via Thresholded Gaussian Processes

Jian Kang
University of Michigan
E-mail: jiankang@umich.edu

Abstract: Blind source separation (BSS) is the separation of latent source signals from their mixtures, which can be achieved by many methods based with different assumptions, criteria or aims, such as principal components analysis (PCA), singular value decomposition (SVD) and independent component analysis (ICA). However, for neuroimaging data analysis, the most existing BSS methods fail to directly account for the spatial dependence among voxels and do not explicitly model the sparsity of source signals. To address those limitations, we propose a Bayesian nonparametric model for BSS of spatial processes. We assume the observed images as the linear mixtures of multiple sparse and piecewise-smooth latent source processes, for which we construct a new class of prior models by thresholding Gaussian processes. We adopt the von-Mises Fisher distribution as the prior model for mixing coefficients. Under some regularity conditions, we show that the proposed model enjoys large prior support; and we establish the consistency of the posterior distribution with a divergent number of voxels in images. The simulation studies demonstrate that the proposed method outperforms the existing ICA methods for latent brain network separation and brain activation region detection. We apply the proposed method to analysis of the resting-state fMRI data in the Kirby 21 dataset, which shows a very promising recovery of latent brain functional networks.

High-Dimensional Spatial Quantile Function-on-Scalar Regression in Neuroimaging Analysis

Linglong Kong
University of Alberta
E-mail: lkong@ualberta.ca

Abstract: In this talk, we develop a novel spatial quantile function-on-scalar regression model, which studies the conditional spatial distribution of a high-dimensional functional response given scalar predictors. With the strength of both quantile regression and copula modeling, we are able to explicitly characterize the conditional distribution of the functional or image response on the whole spatial domain. Our method provides a comprehensive understanding of the effect of scalar covariates at different quantile levels and also gives a practical way to generate new images for given covariate values. Theoretically, we establish the minimax rates of convergence for estimating coefficient functions under both fixed and random designs. We further develop an efficient primal-dual algorithm to handle high-dimensional image data. Simulations and real neuroimaging data analysis are conducted to examine the finite-sample performance. Joint work with Zhengwu Zhang, Xiao Wang and Hongtu Zhu.

A Bayesian State-Space Approach to Mapping Directional Brain Networks

TingTing Zhang
The University of Virginia
E-mail: tz3b@virginia.edu

Abstract: The human brain is a directional network system of brain regions involving directional connectivity. Seizures are a directional network phenomenon as abnormal neuronal activities start from a seizure onset zone (SOZ) and propagate to otherwise healthy brain regions. To localize the SOZ of a patient with drug-resistant epilepsy, clinicians use intracranial EEG (iEEG) to record the patient's neuronal activity inside the skull. iEEG data are high-dimensional multivariate time series recordings of neuronal activities in many small brain regions. We build a state-space multivariate autoregression (SSMAR) for iEEG data to model the underlying directional brain network system. We identify connected brain regions (i.e., mapping the brain network) through estimating the SSMAR parameters that denote directional connectivity. To increase model estimation efficiency and to produce scientifically interpretable network results, we incorporate into SSMAR the scientific knowledge that the underlying brain network tends to have a cluster structure. Specifically, we assign to the SSMAR parameters a stochastic-blockmodel-motivated prior reflecting the cluster structure. In contrast to most existing network models that were developed mainly for observed network edges, we develop a Bayesian framework to estimate the proposed high-dimensional model, infer directional connections, and

identify clusters for the unobserved network edges. We show through both simulation and real data analysis that the new method is robust to various deviations from the model assumptions and outperforms existing network methods. Applying the developed SSMAR and Bayesian approach to an epileptic patient's iEEG data, we reveal the patient's network changes during seizure development and the unique connectivity property of the SOZ.

S149: Designs of Modern Clinical Trials

Integrative analysis of high dimensional data under privacy constraints

Molei Liu
Harvard School of Public Health
E-mail: molei_liu@g.harvard.edu

Abstract: Integrative analysis of high–dimensional data from multiple heterogeneous studies is known to be challenging. The challenge is even more pronounced under the DataSHIELD constraint, under which the individual level data cannot be transferred from the distributed data computers (DC) to the central analysis computer (AC), due to privacy concerns and the summary statistics are shared instead. To overcome this difficulty, we propose a novel framework for high dimensional integrative analysis with generalized linear model that relies solely on the summary data. Performance of our proposed method in estimation, variable selection and simultaneous inference is studied and compared with the ideal individual level data analysis theoretically and numerically. Our approach is shown to be equivalent with the individual level data analysis and dominates existing one-shot approach. Practically, our method facilitates meta-analysis and collaboration in fields like EHR data analysis and GWAS, to improve the statistical power.

Trial Designs for Evaluating Combination HIV Prevention Approaches

Ying Qing Chen
Fred Hutchinson Cancer Research Center
E-mail: yqchen.scharp@gmail.com

Abstract: Recent development of both biomedical and behavioral interventions provides the potential of maximize their population impact in risk reduction of HIV transmission via combination prevention intervention approaches. However, developing powerful and easy-to-implement clinical trial design(s) to assess the effectiveness of combined biomedical and behavioral interventions has been inadequate. We conduct Monte-Carlo simulation studies via the Cox proportional hazards models for time to incident HIV-transmissions to investigate four leading candidate trial designs: 1) single-factor design, 2) factorial design, 3) actives-versus-control "multi-arm" design, and 4) all-versus-none "kitchen-sink" design, for assessing combination prevention intervention approaches. Their potential public health impact is also investigated. In this paper, we compare the pros and cons among the four designs, and argue that the factorial design is an efficient design particularly suitable for combination prevention intervention approaches when multiple candidate interventions are included.

Lessons Learned from Adaptive RCT Designs

Daniel Gillen
University of California
E-mail: dgillen@uci.edu

Abstract: This talk will consider the current state of adaptive clinical trials designs from the perspective of the 2018 FDA guidance document on adaptive designs for drugs and biologics. We will explore the types of adaption that have been proposed and used in regulatory settings, consider the potential efficiency benefits of sample size adaptation, and discuss some of the hard lessons learned from adaptive randomization schemes. The talk will conclude with particular issues that arise when utilizing adaptive designs in the setting of time-to-event endpoints and thoughts on where we go from here in order to increase efficiency and maximize individual- and population-level ethics in randomized controlled trials.

Group Sequential Analysis based on RMST

Lu Tian
Stanford University
E-mail: lutian@stanford.edu

Abstract: It is appealing to coIt is appealing to compared survival analysis based on restricted mean survival time (RMST), since it generates a clinically interpretable summary of the treatment effect, which can be estimated nonparametrically without assuming restrictive assumptions such as the proportional hazards assumption. However, there are special challenges in designing and analyzing group sequential study based on RMST, because that the truncation timepoint of the RMST in the interim analysis often differs from that in the final analysis. A valid test controls the unconditional type one error has been developed in the past. However, there is no appropriate statistical procedure for constructing the confidence interval for the treatment effect measured by the contrast in RMST, while it is crucial for informative clinical decision making. In this talk, we will discuss how to construct confidence intervals for the difference RMST in a group sequential setting. Examples and numerical studies will be presented to illustrate the method. Nduct.

S150: Causal inference and related methodology in health sciences

A Bayesian semiparametric latent variable approach to causal mediation

Yisheng Li
The University of Texas MD Anderson Cancer Center
E-mail: ysli@mdanderson.org

Abstract: In assessing causal mediation effects in randomized studies, a challenge is that the direct and indirect effects can vary across participants due to different measured and unmeasured characteristics. In that case, the population effect estimated from standard approaches implicitly averages over and does not estimate the heterogeneous direct and indirect effects. We propose a Bayesian semiparametric method to estimate heterogeneous direct and indirect effects via clusters, where the clusters are formed by both individual covariate profiles and individual effects due to unmeasured characteristics. These cluster-specific direct and indirect effects can be estimated through a set of regression models where specific coefficients are clustered by a stick-breaking prior. To let clustering be appropriately informed by individual direct and indirect effects, we specify a data-dependent prior. We conduct simulation studies to assess performance of the proposed method compared to other methods.We use this approach to estimate heterogeneous causal direct and indirect effects of an expressive writing intervention for patients with renal cell carcinoma.

Challenge and promise of observational studies in cancer research

Yu Shen

The University of Texas MD Anderson Cancer Center
E-mail: yshen@mdanderson.org

Abstract: An alternative source of data from randomized controlled trials can be found in the large observational databases and longitudinally-followed patient cohorts that have emerged. These invaluable resources present new opportunities in research to provide potential insights into cancer treatment and patient care. However, such studies are not without their own set of challenges. The complexity of sampling mechanisms and various biases associated with prospective observational studies raise considerable analytical challenges in both the design and the data analysis. The peril of selection bias is exacerbated in many cohort studies. To address the above challenges, we need practical statistical designs and innovative analytic approaches to evaluate clinical effectiveness and healthcare interventions outside of controlled clinical trials. We will give an overview on recent semiparametric modeling for right-censored survival data under length-biased sampling. The methods will be reviewed for commonly used proportional hazards model, and AFT model for time-to-event outcomes, and restricted mean survival times. Some related software for the implementation of such methods will be illustrated.

Bidirectional mediation to quantify direct and indirect effects with application to Obesity and Diabetes

Rajesh Talluri
The University of Mississippi Medical Center
E-mail: rajeshstat@gmail.com

Abstract: Obesity and type 2 diabetes are major public health issues with known interdependence. Genetic variants have been associated with obesity, type 2 diabetes, or both; thus, we hypothesize that some single nucleotide polymorphisms (SNPs) associated with both conditions may be mediated through obesity to affect type 2 diabetes or vice versa. We propose a framework for bidirectional mediation analyses. Simulations show that this approach accurately estimates the parameters, whether the mediation is unidirectional or bidirectional. In many scenarios, when the mediator is regressed on the initial variable and the outcome is regressed on the mediator and the initial variable, the resulting residuals are correlated because of other unmeasured covariates not in the model. We show that the proposed model provides accurate estimates in this scenario, too. We applied the proposed approach to investigate the mediating effects of SNPs associated with type 2 diabetes and obesity using genetic data from the Multi-Ethnic Study of Atherosclerosis cohort. Specifically, we used body mass index as a measure for obesity and fasting glucose as a measure for type 2 diabetes. We evaluated the top 6 SNPs associated with both body mass index and fasting glucose. Two SNPs (rs3752355 and rs6087982) had indirect effects on body mass index mediated through fasting glucose (0.2677; 95% confidence interval (CI) [0.0007, 0.6548] and 0.3301; 95% CI [0.0881, 0.8544], respectively). The remaining four SNPs (rs7969190, rs4869710, rs10201400 and rs12421620) directly affect body mass index and fasting glucose without mediating effects.

S151: The Use of Spectral Methods in Statistics: Theory and Applications

The Phase II data for the network of statisticians

Jiashun Jin
Carnegie Mellon University
E-mail: jiashun@stat.cmu.edu

Abstract: We have collected a data set for the networks of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of more than 80K papers published in representative journals in 36 journals in statistics and related fields, spanning about 41 years. The data set provides a fertile ground for research in social networks, text mining, and knowledge discovery, and motivates an array of interesting problem in statistics and machine learning. In this talk, we discuss several problems including overall productivity of statisticians, journal ranking, journal clustering, citation patterns, citation prediction, co-authorship network communities, co-authorship network mixed-memberships, dynamic networks, topic estimation, and dynamic topic estimation. Our analysis uses an array of new methods in network analysis, topic estimation, and neural networks.

Asymptotics of empirical eigenstructure for high dimensional spiked covariance

Weichen Wang
Two Sigma Investments
E-mail: nickweichwang@gmail.com

Abstract: We derive the asymptotic distributions of the spiked eigenvalues and eigenvectors under a generalized and unified asymptotic regime, which takes into account the magnitude of spiked eigenvalues, sample size and dimensionality. This regime allows high dimensionality and diverging eigenvalues and provides new insights into the roles that the leading eigenvalues, sample size and dimensionality play in principal component analysis. Our results are a natural extension of those in [Statist. Sinica 17 (2007) 1617–1642] to a more general setting and solve the rates of convergence problems in [Statist. Sinica 26 (2016) 1747–1770]. They also reveal the biases of estimating leading eigenvalues and eigenvectors by using principal component analysis, and lead to a new covariance estimator for the approximate factor model, called Shrinkage Principal Orthogonal complEment Thresholding (S-POET), that corrects the biases. Our results are successfully applied to outstanding problems in estimation of risks for large portfolios and false discovery proportions for dependent test statistics and are illustrated by simulation studies.

Detecting Rare and Weak Spikes in Large Covariance Matrices

Tracy Ke
Harvard University
E-mail: zke@fas.harvard.edu

Abstract: Given p-dimensional Gaussian vectors X1, ..., Xn, where $p \geq n$, we are interested in testing a null hypothesis where $\Sigma = Ip$ against an alternative hypothesis where all eigenvalues of Σ are 1, except for r of them are larger than 1 (i.e., spiked eigenvalues). We consider a Rare/Weak setting where the spikes are sparse (i.e., $1<<r<<p$) and individually weak (i.e., each spiked eigenvalue is only slightly larger than 1), and discover a phase transition: the two-dimensional phase space that calibrates the spike sparsity and strengths partitions into the Region of Impossibility and the Region of Possibility. In Region of Impossibility, all tests are (asymptotically) powerless in separating the alternative from the null. In Region of Possibility, there are tests that have (asymptotically) full power. We consider a CuSum test, a trace-based test, an eigenvalue-based Higher Criticism test, and a Tracy-Widom test, and show that the first two tests have asymptotically full power in Region of Possibility. To use our results from a different angle, we derive new bounds for (a) empirical eigenvalues,

and (b) cumulative sums of the empirical eigenvalues, both under the alternative hypothesis. Part (a) is related to those in the literature, but both the settings and results are different. The study requires careful analysis of the L1-distance of our testing problem and delicate Radom Matrix Theory. Our technical devises include (a) a Gaussian proxy model, (b) Le Cam's comparison of experiments, and (c) large deviation bounds on empirical eigenvalues.

A geometric perspective of hypothesis testing

Yuting Wei
Carnegie Mellon University
E-mail: ytwei@cmu.edu

Abstract: This talk is devoted to understanding the behavior of compound testing problems within the Gaussian sequence model from a geometric perspective. In this talk, two vignettes are considered.

When the null and alternative are specified by a pair of closed, convex cones---the case arising in various applications, including detection of treatment effects, trend detection in econometrics, signal detection in radar processing, and shape-constrained inference in non-parametric statistics, we studied the behavior of the generalized likelihood ratio test (GLRT). Despite the wide-spread use of the GLRT, its properties have yet to be fully understood. When is it optimal, and when can it be improved upon? How does its performance depend on the cones? I provide some answers to these and other questions, all based on a tight characterization of the GLRT's performance.

When the hypotheses are given by a known vector within a high dimensional ellipse, or other unknown vectors in the ellipse---the case underlying the heart of non-parametric goodness-of-fit testing, signal detection in cognitive radio, and regression function testing in reproducing kernel Hilbert spaces, we study difficulty in a way that is adaptive to vectors within the ellipse. By characterizing the localized minimax testing radius sharply, our results yield interesting phenomena that were not known before.

S152: High dimensional analysis and application in biomarker identification

Independence Structure Test in Ultra High-Dimensional Data

Jing He
Southwestern University of Finance and Economics
E-mail: he_jing@swufe.edu.cn

Abstract: This paper considers testing for independence structure in ultra high dimensional data. This problem includes testing of mutual independence between components of a random vector as a special case. The dependence between two random vectors can be characterized by the projection correlation proposed by Zhu et al. (2017), which has many appealing properties. To test for the independence structure of a random vector X, we propose to use the maximum projection correlations between the subvectors of X as the test statistic. It can be modified as the maximum sum-of-squared type test statistic to increase the power for against the dense alternative. Considering that it is difficult to calculate the V-statistic estimator of the projection correlation when n and p is extremely large, we propose the blockwise computation method, which is conductive to parallel computing. Simulation study shows that the proposed method can greatly reduce the overall computation time. We further demonstrate the performances of the proposed test in simulations, as well as an empirical study on a real application dataset.

Prognostic biomarker identification and subgroup analysis using high dimensional inference in CAR-T cell immunotherapy trial

Qian Wu
Fred Hutch
E-mail: wuqian7@gmail.com

Abstract: Chimeric antigen receptor (CAR) T cell immunotherapy has shown remarkable efficacy in patients with relapsed CD19+ B-cell non-Hodgkin lymphoma (NHL). Durable response has been observed in a subset of patients with lymphodepletion chemotherapy followed by infusion of T cells. One of the important objectives is to identify patients' subgroups defined by one or multiple factors associated with progression-free-survival (PFS) and overall survival (OS), though most of the immunotherapy trials are still in the stage of phase I/II with limited sample size but large number of covariates, such as cytokine and manufactory biomarkers. Univariate Cox regression showed many covariates with marginal significance and they are highly correlated. Traditional variables selection approach, such as forward/backward/stepwise regression, cannot deal with large p and small n (p>>n) and might not be good to select variables highly correlated. Penalized regression, such as LASSO and Elastic Net, can deal with p>>n and address multicollinearity, though output provides biased coefficient without p-value and confidence interval, where we cannot control false discovery rate with high-dimensional data. We also observed un-stable results by penalized regression with small sample size n (~50) but large number of covariates p (p>100). Recently, high dimensional inference (Fang et al., 2017) proposed de-correlated approach for Cox regression (CoxHDI), which is flexible to choose different penalized regression, stable for the choice of tuning parameter, provide de-biased coefficient, p-value and confidence interval. We proposed a stability approach (freqNet) based on Elastic Net but run 100 times and selected three biomarkers (pre-lymphodepletion serum LDH, MCP-1, IL-7) with frequency > 85% (Hirayama et al., 2019). Extensive simulation is conducting to compare the performance between CoxHDI and freqNet, and we demonstrate that the low dimensional empirical type I error rate is controlled when p >> n, though observed type I error inflation when p << n for both methods. We further applied CoxHDI on the same NHL dataset, and observed the same three biomarkers with p-value < 0.05 and one more biomarker (TNFRp75) identified, which also have good biological explanation associated with PFS.

Threshold-based subgroup testing in logistic regression models

Ying Huang
Fred Hutchinson Cancer Research Center
E-mail: yhuang124@gmail.com

Abstract: Associations between disease and predictors can differ across subgroups characterized by other covariates including treatment. In this paper, we consider a hypothesis testing problem for the existence of subgroups with heterogeneous disease risk models. Allowing multivariate predictors and/or covariates, we develop inferential procedures based on maximum of likelihood-ratio statistics in a threshold-based framework, where subgroups are characterized by unknown linear combinations of covariates. Numerical studies demonstrate the advantage of the proposed method compared to the alternative, existing two-step strategy that separates the estimation of the covariate combination.

From the testing of the threshold effect based on the estimated

combination score. We further extend the method to two-phase sampling settings where the complete set of variables are only available from a subset of participants in the phase one cohort. We demonstrate the application of our method using a real example from a recent HIV vaccine trial, where we test for the existence of subgroups based on the Fc receptor genes that modified vaccine's effects on HIV acquisition risk.

Information Enhanced Model Selection for High-Dimentional Gaussian Graphic Model with Application to Metabolomics Data

Jiang Gui
Dartmouth College
E-mail: jiang.gui@dartmouth.edu

Abstract: In light of the low signal-to-noise nature of many large biological data sets, we propose a novel method to identify the structure of association networks using a Gaussian graphical model combined with prior knowledge. Our algorithm includes the following two parts. In the first part we propose a model selection criterion called structural Bayesian information criterion (SBIC) in which the prior structure is modeled and incorporated into the Bayesian information criterion (BIC). It is shown that the popular extended BIC (EBIC) is a special case of SBIC. In second part we propose a two-step algorithm to construct the candidate model pool. The algorithm is data-driven and the prior structure is embedded into the candidate model automatically. Theoretical investigation shows that under some mild conditions SBIC is a consistent model selection criterion for high-dimensional Gaussian graphical model. Simulation studies validate the superiority of the SBIC over the standard BIC and show the robustness to the model misspecification. Application to untargeted profile data from infant feces collected from subjects enrolled in a large molecular epidemiologic cohort study validates that prior knowledge on metabolic pathway involvement is a significant factor for the conditional dependence among metabolites. More importantly new relationships among metabolites are identified through the proposed algorithm which can not be covered by standard pathway analysis. Some of them have been widely recognized in literature.

S153: Recent Advances in Ultrahigh Dimensional Data

The Lq-norm learning for ultrahigh-dimensional survival data: an integrative framework

Xuerong Chen
Southwestern University of Finance and Economics
E-mail: chenxuerong@swufe.edu.cn

Abstract: In the era of precision medicine, survival outcome data with high-throughput predictors are routinely collected from many biomedical studies. Models with an exceedingly large number of covariates are either infeasible to fit or likely to incur low predictability because of overfitting. Variable screening is key in identifying and removing irrelevant attributes. Recent years have seen a surge in screening methods, but most of them rely on some particular modeling assumptions. Motivated by a study on detecting gene signatures for multiple myeloma patients' survival, we propose a model-free Lq-norm learning procedure, which includes the well-known Cram´er–von Mises and Kolmogorov criteria as two special cases. The work provides an integrative framework for detecting predictors with various levels of impact, such as short- or long-term impact, on censored outcome data. The framework naturally leads to a scheme which combines results from different q0s to reduce false negatives, an aspect often overlooked by the current literature. We show that our method possesses sure screening properties. The utility of the proposal is confirmed with simulation studies and an analysis of the multiple myeloma study.

Testing for Homogeneity of Mean Vectors and Covariance Matrices in High-dimension

Wenwen Guo
School of Mathematical Sciences, Capital Normal University , China
E-mail: guowenwen114@163.com

Abstract: This paper aims to develop new tests for homogenity of mean vectors and covariance matrices in high-dimension. Two categorically weighted measures on the difference of means and covariances are proposed respectively. The asymptotic distributions of the two statistics are explored. We further give simple algorithms to facilitate applications. Finite sample performance of the two tests are shown through simulations.

Ball Covariance: A Generic Measure of Dependence in Banach Space

Wenliang Pan
Sun Yat-sen university
E-mail: panwliang@mail.sysu.edu.cn

Abstract: Technological advances in science and engineering have led to the routine collection of large and complex data objects, where the dependence structure among those objects is often of great interest. Those complex objects (e.g, different brain subcortical structures) often reside in some Banach spaces, and hence their relationship cannot be well characterized by most of the existing measures of dependence such as correlation coefficients developed in Hilbert spaces. To overcome the limitations of the existing measures, we propose Ball Covariance as a generic measure of dependence between two random objects in two possibly different Banach spaces. Our Ball Covariance possesses the following attractive properties: (i) It is nonparametric and model-free, which make the proposed measure robust to model mis-specication; (ii) It is nonnegative and equal to zero if and only if two random objects in two separable Banach spaces are independent; (iii) Empirical Ball Covariance is easy to compute and can be used as a test statistic of independence. We present both theoretical and numerical results to reveal the potential power of the Ball Covariance in detecting dependence. Also importantly, we analyze two real datasets to demonstrate the usefulness of Ball Covariance in the complex dependence detection.

A nonparametric test for proportional covariance matrices in large dimension and small samples

Kai Xu
School of Mathematics and Statistics, Anhui Normal University
E-mail: tjxxukai@163.com

Abstract: This work is concerned with testing the proportionality between two high-dimensional covariance matrices. Several tests for proportional covariance matrices, based on modifying the classical likelihood ratio test and applicable in high dimension, have been proposed in the literature. Despite their usefulness, they tend to have unsatisfactory performance for nonnormal high-dimensional multivariate data in terms of size or power. This article proposes a new high-dimensional test by developing a bias correction to the existing test statistic constructed based on a scaled distance measure. The suggested test is nonparametric without requiring any specific parametric distribution such as the normality assumption. It can

accommodate scenarios where the data dimension p is greater than the sample size n, namely the ``large p, small n" problem.

With the aid of tools in modern probability theory, we study theoretical properties of the newly proposed test, which include the asymptotic normality and a power evaluation.

We demonstrate empirically that our proposal has good size and power performances for a range of dimensions, sample sizes and distributions in comparison with the existing counterparts.

S155: Recent advances in biomedical big data analytics

MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting linkage disequilibrium and horizontal pleiotropy

Jin Liu
Duke-NUS Medical School
E-mail: jin.liu@duke-nus.edu.sg

Abstract: The proliferation of genome-wide association studies (GWAS) has prompted the use of two-sample Mendelian randomization (MR) with genetic variants as instrumental variables (IV) for drawing reliable causal relationships between health risk factors and disease outcomes. However, the unique features of GWAS demand MR methods account for both linkage disequilibrium (LD) and ubiquitously existing horizontal pleiotropy among complex traits, which is a phenomenon that a variant affects the outcome other than exclusively through the exposure. Therefore, statistical methods that fail to consider LD and horizontal pleiotropy can lead to biased estimates and false-positive causal relationships. To overcome these limitations, we propose a probabilistic model for MR analysis to identify casual effect between risk factors and disease outcomes by using GWAS summary statistics in the presence of LD, as well as properly accounts for horizontal Pleiotropy among genetic variants (MR-LDP). MR-LDP utilizes a computationally efficient parameter-expanded variational Bayes expectation-maximization (PX-VBEM) algorithm, calibrating the evidence lower bound (ELBO) for a likelihood ratio test. We further conducted comprehensive simulation studies to demonstrate the advantages of MR-LDP over existing methods in terms of both type-I error control and point estimates. Moreover, we used two real exposure-outcome pairs (CAD-CAD and BMI-BMI; CAD for coronary artery disease and BMI for body mass index) to validate results from MR-LDP in comparison with alternative methods, particularly showing that our method is more efficient using all instrumental variants in LD. By further applying MR-LDP to lipid traits and BMI as risk factors on complex diseases, we identified multiple pairs of significant causal relationships, including protective effect of high-density lipoprotein cholesterol (HDL-C) on peripheral vascular disease (PVD), and positive causal effect of body mass index (BMI) on haemorrhoids.

Model-based microbiome data ordination: A variational approximation approach

Tao Wang
Shanghai Jiao Tong University
E-mail: neowangtao@sjtu.edu.cn

Abstract: The coevolution between human and bacteria colonizing the human body has profound implications for heath and development, with a growing body of evidence linking the altered microbiome composition with a wide array of disease states. Yet dimension reduction and visualization analysis of microbiome data is still in its infancy and many challenges exist. In this talk we introduce a general framework, Zero-Inflated Probabilistic PCA (ZIPPCA, pronounced zipcar), for dimension reduction and data ordination of multivariate abundance data, and propose an efficient variational approximation method for estimation, inference, and prediction. Extensive simulations show that the proposed method outperforms algorithm-based methods and compares favorably with existing model-based methods. We further apply our method to a gut microbiome dataset for visualization analysis of community composition across age and geography.

Post-GWAS data integration identifies risk factors for Alzheimer's disease

Qiongshi Lu
University of Wisconsin-Madison
E-mail: qlu@biostat.wisc.edu

Abstract: Despite the findings in genome-wide association studies (GWAS) for late-onset Alzheimer's disease (LOAD), our understanding of its genetic architecture is far from complete. Transcriptome-wide association analysis that integrates GWAS data with large-scale transcriptomic databases is a powerful method to study the genetic architecture of complex traits. However, it is challenging to effectively utilize transcriptomic information given limited and unbalanced sample sizes in different tissues. Here we introduce and apply UTMOST, a principled framework to jointly impute gene expression across multiple tissues and perform cross-tissue gene-level association analysis using GWAS summary statistics. Compared with single-tissue methods, UTMOST achieved 39% improvement in expression imputation accuracy and generated effective imputation models for 120% more genes in each tissue. A total of 69 genes reached the Bonferroni-corrected significance level in the transcriptome-wide association meta-analysis for LOAD. Among these findings, we identified novel risk genes at known LOAD-associated loci as well as five novel risk loci. Several genes, including IL10 and ADRA1A, also have therapeutic potential to improve neurodegeneration. Cross-tissue conditional analysis further fine-mapped IL10 as the functional gene at the CR1 locus, a well-replicated risk locus for LOAD. Extension of this framework to perform biobank-wide association analysis will also be discussed. Overall, integrated analysis of transcriptomic annotations and biobank information provides insights into the genetic basis of LOAD and may guide functional studies in the future.

Statistical methods for data integration in single-cell genomics

Zhixiang Lin
The Chinese University of Hong Kong
E-mail: zhixianglin@cuhk.edu.hk

Abstract: Some recent work regarding integrative analysis of single-cell genomic data will be presented, including the integration of multiple types of genomic data and the integration of single-cell genomic data with bulk genomic data.

S156: Recent advances in complex biological data modeling

Covariate-dependent graphs with application in cancer genomics

Yang Ni
Texas A&M University
E-mail: yni@stat.tamu.edu

Abstract: We consider the problem of modeling conditional independence

structures in heterogeneous data in the presence of additional subject-level covariates. We propose a novel specification of a conditional (in)dependence function of covariates – which allows the structure of a directed or undirected graph to vary with the covariates and produces both subject-specific and predictive graphs. Applications in cancer genomics will presented.

Integrative analysis of multi-platform data

Jianhua Hu
Columbia University
E-mail: jh3992@cumc.columbia.edu

Abstract: We propose a statistical framework of shared informative factor models that can jointly analyze multi-platform omic data and explore their associations with a disease phenotype. The common disease-associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype. Extensive simulation studies and an application demonstrate the performance of the proposed method in terms of biomarker detection accuracy.

Semiparametric Model for Bivariate Survival Data Subject to Biased Sampling

Jin Piao
University of Southern California
E-mail: jean.pjin@gmail.com

Abstract: To better understand the relationship between patient characteristics and their residual survival after an intermediate event such as the local cancer recurrence, it is of interest to identify patients with the intermediate event and then analyze their residual survival data. One challenge in analyzing such data is that the observed residual survival times tend to be longer than those in the target population, since patients who die before experiencing the intermediate event are excluded from the identified cohort. We propose to jointly model the ordered bivariate survival data using a copula model and appropriately adjusting for the sampling bias.

We develop an estimating procedure to simultaneously estimate the parameters for the marginal survival functions and the association parameter in the copula model, and use a two-stage expectation-maximization algorithm. Using empirical process theory, we prove that the estimators have strong consistency and asymptotic normality. We conduct simulations studies to evaluate the finite sample performance of the proposed method. We apply the proposed method to two cohort studies to evaluate the association between patient characteristics and residual survival.

Generalized probabilistic principal component analysis

Weining Shen
UC Irvine
E-mail: weinings@uci.edu

Abstract: Principal component analysis (PCA) is a well-established tool in machine learning and data processing. The principal axes in PCA were shown to be equivalent to the maximum marginal likelihood estimator of the factor loading matrix in a latent factor model for the observed data, assuming that the latent factors are independently distributed as standard normal distributions. However, the independence assumption may be unrealistic for many scenarios such as modeling multiple time series, spatial processes, and functional data, where the outcomes are correlated. In this paper, we introduce the generalized probabilistic principal component analysis (GPPCA) to study the latent factor model for multiple correlated outcomes, where each factor is modeled by a Gaussian process. Our method generalizes the previous probabilistic formulation of PCA (PPCA) by providing the closed-form maximum marginal likelihood estimator of the factor loadings and other parameters. Based on the explicit expression of the precision matrix in the marginal likelihood that we derived, the number of the computational operations is linear to the number of output variables. Furthermore, we also provide the closed-form expression of the marginal likelihood when other covariates are included in the mean structure. We highlight the advantage of GPPCA in terms of the practical relevance, estimation accuracy and computational convenience. Numerical studies of simulated and real data confirm the excellent finite-sample performance of the proposed approach.

S157: Deep learning and applications

Invariant Data Representations with Multiscale Mathematical Models for ConvNets

Matthew Hirn
Michigan State University
E-mail: mhirn@msu.edu

Abstract: Convolutional neural networks (ConvNets) have revolutionized our approach to learning tasks for high dimensional signal data by processing a signal through a cascade of learned convolution operators and nonlinear operations, which successively extract information from the signal that can be used for downstream tasks such as classification or regression. Motivated by the successes of ConvNets, in this talk I will introduce the wavelet scattering transform, which can be viewed as a simplified mathematical model for them. This transform replaces the learned filters of ConvNets with predefined wavelets, which are multiscale, oscillating waveforms with zero average, and computes a cascade of alternating wavelet transforms and nonlinear operators. Unlike ConvNets, which are task driven, a scattering transform is motivated by invariance and stability properties inherent in the data. Here we will focus on problems at the interface of invariant representation learning and statistics, such as texture classification and synthesis, parameter estimation of random processes, energy prediction of amorphous solids, and multi-reference alignment inverse problems. We will show that invariant data measurements derived from the wavelet scattering transform can be a useful tool in tackling these problems.

Robustness and Sparsity in Deep Learning

Yuan Yao
Hong Kong University of Science and Technology
mail: yuany@ust.hk

Abstract: We are going to talk about achieving robustness from generative adversarial networks and learning structural sparsity via differential inclusion method.

PDE-based Methods for Interpolation on High Dimensional Point Cloud

Zuoqiang Shi
Tsinghua University
E-mail: zqshi@tsinghua.edu.cn

Abstract: Interpolation on high dimensional point cloud provides a fundamental model in many data analysis and machine learning problems. In this talk, we will present some PDE based methods to do interpolation on

point cloud. Applications in image processing and machine learning are shown to demonstrate the performance of our methods.

Predicting Plant Stress Responses Using Deep Neural Network

Yuying Xie
Michigan State University
E-mail: xyy@msu.edu

Abstract: Plants exhibit diverse responses under environmental stresses through regulating the expression of related genes. The ability to predict the genic responses to abiotic stresses using genomic sequence data will yield a deeper understanding on how plants will respond to environmental challenges and hence facilitate crop breeding for extreme environments which is increasingly common due to climate change. Deep learning-based prediction model has been successfully applied to numerous real-world applications including bioinformatics due to its ability to extract high-level features from massive data. However, building a successful and efficient deep neural network model requires access to a massive amount of data, which is rare in most of the plant species. We developed an effective deep transfer learning framework to predict genic response to abiotic stress, leveraging data from different species and different stresses. We applied this model in arabidopsis dataset and outperformed other existing models in term of prediction accuracy.

S158: Modern Clinical Trial Design and Analysis Methods

Adaptive borrowing of information across patient subgroups in a basket trial based on distributional discrepancy

Haiyan Zheng
Newcastle University
E-mail: haiyan.zheng@newcastle.ac.uk

Abstract: Basket trials emerge as a new class of efficient approaches to evaluate a treatment in several patient subgroups simultaneously. In this paper, we develop a novel analysis methodology for early phase basket trials, which enables borrowing of information to improve decision making in a collection of subgroups. For each subgroup-specific parameter that underpins the treatment effect, a marginal predictive prior (MPP) is specified using information from the other subgroups. More specifically, it comprises a series of commensurate predictive priors (CPPs), each with a normal precision that captures the commensurability of information. We propose using a distributional discrepancy to characterise the pairwise commensurability between any two subgroups, so as to inform the choice of a spike-and-slab prior to be placed on the normal precision. This determines the degree of borrowing from an external subgroup. When there exist at least three subgroups in a basket trial, we convert the pairwise discrepancy measurements into a set of normalised weights and allocate them to the CPPs accordingly. This leads to an MPP that leverages only information from the most consistent subgroups. The MPP is then updated using the contemporary subgroup data to a robust posterior. Trial operating characteristics of the proposed methodology are evaluated through simulations motivated by a real clinical trial. Compared with alternative Bayesian analysis models, our proposal is more advantageous for (i) identifying the most consistent subgroups, and (ii) gauging the amount of information to be leveraged. Numerical results also suggest that our analysis methodology can potentially improve the precision of estimates and the statistical power for hypothesis testing.

Estimation of survival under dependent truncation

Jing Qian
University of Massachusetts
E-mail: qian@schoolph.umass.edu

Abstract: In clinical trials designed to assess the treatment effect on survival, patients are required to survive from the time of diagnosis to recruitment. This results in samples with left truncated distributions of event time. Standard survival analysis methods for estimation of the distribution of the event time require quasi-independence of event time and truncation time. When quasi-independence does not hold, standard methods may yield biased estimation. We propose two types of methods for estimation of survival under dependent truncation. One is a transformation model approach to model a latent quasi-independent truncation time as a function of the observed dependent truncation time and the event time and an unknown transformation parameter. The proposed method can accommodate right censored data. The other one is an inverse-probability-weighting type approach to accommodate the dependent truncation induced by covariates. We evaluate the proposed methods through extensive simulations and apply them to clinical studies of neurological diseases.

How to apply the multilevel modeling in large health care administrative data

Jun Guan
Methodologist
E-mail: jun.guan@ices.on.ca

Abstract: In health care research, we often need to analyze the hierarchical or multi-level, or nested data. For example, long-term care residents living in the same nursing home have access to similar physicians and nurses with similar prescribing practices. Patients operated on by the same surgeon have outcomes that are more similar than those operated on by a different surgeon since outcome partly depends on the surgeon's skill; furthermore, patients of surgeons working at the same hospitals may have similar outcomes because post-operative care varies among hospitals. In this talk, I will share some experience on how the multilevel modeling has been applied in large health administrative data, plan to address some challenges when we are analyzing the health care big data (millions records), and demonstrate some tips/tricks on how we have handled multilevel modeling in some software such as SAS, R, STATA.

Keywords: multilevel modeling, health administrative data, SAS, R, STATA.

S159: Recnet Develpments in Statistical Network Analysis

Hierarchical community detection by recursive partitioning

Tianxi Li
University of Virginia
E-mail: tianxili@virginia.edu

Abstract: The problem of community detection in networks is usually formulated as finding a single partition of the network into some "correct" number of communities. We argue that it is more interpretable and in some regimes more accurate to construct a hierarchical tree of communities instead. This can be done with a simple top-down recursive partitioning algorithm, starting with a single community and separating the nodes into two communities by spectral clustering repeatedly, until a stopping rule suggests there are no further communities. This class of algorithms is model-free, computationally efficient, and requires no tuning other than

selecting a stopping rule. We show that there are regimes where this approach outperforms K-way spectral clustering, and propose a natural framework for analyzing the algorithm's theoretical performance, the binary tree stochastic block model. Under this model, we prove that the algorithm correctly recovers the entire community tree under relatively mild assumptions. We also apply the algorithm to a dataset of statistics papers to construct a hierarchical tree of statistical research communities.

Popularity-Adjusted Block Models for Networks with Community Structure

Yuguo Chen
University of Illinois at Urbana-Champaign
E-mail: yuguo@illinois.edu

Abstract: The community structure observed in empirical networks has been of particular interest in the statistics literature, with a strong emphasis on the study of block models. We study an important network feature called node popularity, which is closely associated with community structure. Neither the classical stochastic block model nor its degree-corrected extension can satisfactorily capture the dynamics of node popularity as observed in empirical networks. We propose a popularity-adjusted block model for flexible and realistic modeling of node popularity. We establish consistency of likelihood modularity for community detection as well as estimation of node popularities and model parameters, and demonstrate the advantages of the new modularity over the degree-corrected block model modularity in simulations. By analyzing the political blogs network, the British MP network, and the DBLP bibliographical network, we illustrate that improved empirical insights can be gained through this methodology.

Network Differential Connectivity Analysis

Ali Shojaie
University of Washington
E-mail: ashojaie@uw.edu

Abstract: Recent evidence suggests that changes in biological networks, e.g., rewiring or disruption of key interactions, may be associated with development of complex diseases. These findings have motivated new research initiatives in computational and experimental biology that aim to obtain condition-specific estimates of biological networks, e.g. for normal and tumor samples, and identify differential patterns of connectivity in such networks, known as differential network biology.

In this talk, we focus on testing whether two Gaussian graphical models are the same. Existing methods try to accomplish this goal by either directly comparing their estimated structures, or testing the null hypothesis that the partial correlation values are equal. Unfortunately, these methods may lead to misleading results. To address this shortcoming, we propose a two-step inference framework, for testing the null hypothesis that the edge sets in two networks are the same. The proposed framework is especially appropriate if the goal is to identify nodes or edges that show differential connectivity. We investigate theoretical and numerical properties of the proposed framework and illustrate its utility in a study of changes in brain connectivity network associated with mild trauma.

Edgeworth approximation to network U-statistics

Yuan Zhang
Ohio State University
E-mail: yzhanghf@stat.osu.edu

Abstract: In this talk, we present some limiting theory and numerical results for approximating the distribution of network U-statistics in presence of edge-wise random noise. We compare the results with the classical noiseless setting and discuss their connections and differences. We derive the 1/sqrt(n) order term in the Edgeworth approximation and show that the accuracy of such approximation achieves order 1/n, ignoring the logarithm factor, under mild conditions. We also discuss the nonparametric bootstrap and compared our method's accuracy with it.

S160: Causal Inference

Estimation of Optimal Individualized Treatment Rule Using the Covariate-Specific Treatment Effect Curve with High-dimensional Covariates

Xiaohua Zhou
Beijing International Center for Mathematical Research and Department of Biostatistics, Peking University
E-mail: azhou@math.pku.edu.cn

Abstract: In this talk, we propose a new semi-parametric modeling strategy for heterogeneous treatment effect estimation and individualized treatment selection with a large number of baseline covariates. To achieve our goals, we first extend the concept of a covariate-specific treatment effect (CSTE) curve originally proposed by Zhou and Ma (2013) to the situation with high-dimensional covariates. The CSTE curve is estimated by a spline-backfitted kernel procedure, which enables us to further construct a simultaneous confidence band (SCB) for the CSTE curve under a desired confidence level. Based on the SCB, we then find the subgroups of patients that benefit from each treatment, so that we can make individualized treatment selection. The innovations of the proposed method are three-fold. First, the proposed method can quantify variability associated with the estimated optimal individualized treatment rule with high-dimensional covariates. Second, the proposed method is very flexible to depict both local and global associations between the treatment and baseline co-variates in the presence of high-dimensional covariates, and thus is robust against model mis-specification. Third, the proposed method enjoys some good theoretical properties and hence can provide a sound basis for conducting statistical inference in making individualized treatment decisions with high-dimensional covariates. This is a joint wok with Wenchuan Guo at Bristol-Myers Squibb and Shujie Ma at University of California Riverside.

Specification tests for generalized propensity scores using double projections

Xiaojun Song
Peking University
E-mail: sxj@gsm.pku.edu.cn

Abstract: In this article we propose a new class of nonparametric tests for testing the correction specification of generalized propensity score models based on double projections that is suitable particularly with high-dimensional covariates. The first projection has been introduced in the literature to address the high dimensionality of available covariates and to apply a specification test to one-dimensional projections of the covariates, while the second projection is particularly useful for eliminating the parameter estimation effect and also facilitates a convenient multiplier bootstrap procedure to implement the proposed tests. The combination of two projections, termed as double projections, delivers a nice diagnostics tool that is easy to use in practice and powerful against a broad class of alternatives even in the presence of high-dimensional covariates.

Covariate Adjustment in Completely Randomized Experiments

With Noncompliance

Hanzhong Liu
Tsinghua University
E-mail: lhz2016@mail.tsinghua.edu.cn

Abstract: Noncompliance is a common problem in completely randomized experiments. When there is noncompliance, investigators are often interested in estimating the complier average causal effect (CACE) and regression adjustments are often used to gain estimation efficiency. The indirect least squares (ILS) and the two stage least squares (TSLS) are two commonly used regression adjustment methods. We show that they are numerically equal to each other under the Neyman-Rubin potential outcomes framework. They are asymptotically normal with the true CACE as their mean, but their asymptotic variances may be larger than that of the unadjusted Wald estimator. In order to reduce the variance, we propose to include the covariates by treatment assignment interaction in the ILS, as in the case of estimating average causal inference without noncompliance. Under mild conditions, we show that this estimator is consistent and asymptotically normal with asymptotic variance no greater than that of the Wald estimator even when the number of covariates is larger than the number of observations. We provide a conservative estimator of the asymptotic variance, which can yield tighter confidence intervals than the Wald estimator. Moreover, we study the TSLS with covariates by treatment received interaction added in the model and show that it can be asymptotically biased if the covariates are not centered appropriately. Simulation studies show that ILS with interaction can be advantageous when compared with other methods.

On the Efficiency of Logistic Regression Estimators in Estimating The Causal Effect

Jinzhu Jia
Peking University
E-mail: jzjia@math.pku.edu.cn

Abstract: This talk is about the Logistic Regression Estimators of the Average Treatment Effect (ATE) in randomized experiments, when the potential outcomes are binary variables. Based on some regularity conditions, exact expressions of asymptotic variances of the estimators are provided which is considered to be an important criterion for evaluating asymptotic efficiency. Also, we'll compare efficiency of different estimators of ATE (including the regression adjustment estimator) and prove some important results. We find that introducing interaction terms between the assignment variable and the covariate sometimes helps improve asymptotic efficiency. Numerical simulations are carried out to verify the theoretical results and a counterexample is given to show that sometimes introducing interaction terms in Logistic Regression might make the estimator of ATE less efficient.

S161: Recent Advances in Statistical Learning for Healthcare and Biomedical Problems

Minorization-Maximization-based Boosting for Large-scale Survival Analysis with Time-Varying Effects

Zhi (Kevin) He
University of Michigan
E-mail: kevinhe@umich.edu

Abstract: National disease registries have produced a vast amount of data. Many existing statistical methods that perform well for moderate sample sizes and small-dimensional data do not scale to such large-scale data, leading to a demand for statistical techniques that enable full utilization of these rich sources of information. For example, the time-varying effects model is a flexible and powerful tool for modeling the dynamic changes of covariate effects. However, in survival analysis, its computational burden increases quickly as the number of sample sizes or predictors grows. Traditional methods that perform well for moderate sample sizes and low-dimensional data do not scale to massive data. Analysis of national kidney transplant data with a massive sample size and large number of predictors defy any existing statistical methods and software. In view of these difficulties, we propose a Minorization-Maximization-based boosting procedure for estimating the time-varying effects. Leveraging the block structure formed by the basis expansions, the proposed procedure iteratively updates the optimal block-wise direction along which the approximate increase in the log-partial likelihood is maximized. The resulting estimates ensure the ascent property and serve as refinements of the previous step. The performance of the proposed method is examined by simulations and applications to the analysis of national kidney transplant data.

Functional Regression for Brain Imaging

Bin Nan
University of California, Irvine
E-mail: nanb@uci.edu

Abstract: It is well-known that the major challenges in analyzing imaging data arise from spatial correlation and high-dimensionality of voxels. Our primary motivation and application come from brain imaging studies on cognitive impairment in elderly subjects with brain disorders. We propose an efficient regularized Haar wavelet-based approach for the analysis of three-dimensional brain image data in the framework of functional data analysis, which automatically takes into account the spatial information among neighboring voxels. We conduct extensive simulation studies to evaluate the prediction performance of the proposed approach and its ability to identify related regions to response variable, with the underlying assumption that only few relatively small subregions are associated with the response variable. We then apply the proposed method to searching for brain subregions that are associated with cognition using PET images of patients with Alzheimer's disease, patients with mild cognitive impairment, and normal controls. Additional challenges, current and future directions of statistical methods in imaging analysis of AD will also be discussed.

A Bayesian Approach to Joint Estimation of Multiple Graphical Models

George Michailidis
U of Florida
E-mail: gmichail@ufl.edu

Abstract: The problem of joint estimation of multiple graphical models from high dimensional data has been studied in the statistics and machine learning literature, due to its importance in diverse fields including molecular biology, neuroscience and the social sciences. This work develops a Bayesian approach that decomposes the model parameters across the multiple graphical models into shared components across subsets of models and edges, and idiosyncratic ones. Further, it leverages a novel multivariate prior distribution, coupled with a pseudo-likelihood that enables fast computations through a robust and efficient Gibbs sampling scheme. We establish strong posterior consistency for model selection, as well as estimation of model parameters under high dimensional scaling with the number of variables growing exponentially with the sample size. The

efficacy of the proposed approach is illustrated on both synthetic and real data.

S162: Statistical models for diseases with spatial or temporal variations

Modeling heroin-related EMS calls in space and time

Zehang Li
Yale School of Public Health
E-mail: lizehang@gmail.com

Abstract: Opioid use and overdose have become an important public health issues in the United States. However, understanding the spatial and temporal dynamics of opioid overdose incidents and effects of public health interventions and policy changes can be challenging. Effects may be heterogeneous across space and time, and may exhibit spillovers to regions in which the intervention did not take place. In this talk, we discuss considerations in mapping the risk of overdose in small areas over time, and models to characterize the dynamics of overdose incidents. We also outline a framework for estimating causal impacts of public health interventions from surveillance data under spatial-temporal confounding.

An ensemble approach to predicting the impact of vaccination on rotavirus disease in Niger

Jaewoo Park
Yonsei University
E-mail: jaewoopark1201@gmail.com

Abstract: Recently developed vaccines provide a new way of controlling rotavirus in sub-Saharan Africa. Models for the transmission dynamics of rotavirus are critical both for estimating current burden from imperfect surveillance and for assessing potential effects of vaccine intervention strategies. We examine rotavirus infection in the Maradi area in southern Niger using hospital surveillance data provided by Epicentre collected over two years. Additionally, a cluster survey of households in the region allows us to estimate the proportion of children with diarrhea who consulted at a health structure. Model fit and future projections are necessarily particular to a given model; thus, where there are competing models for the underlying epidemiology an ensemble approach can account for that uncertainty. We compare our results across several variants of Susceptible-Infectious-Recovered (SIR) compartmental models to quantify the impact of modeling assumptions on our estimates. Model-specific parameters are estimated by Bayesian inference using Markov chain Monte Carlo. We then use Bayesian model averaging to generate ensemble estimates of the current dynamics, including estimates of R0, the burden of infection in the region, as well as the impact of vaccination on both the short-term dynamics and the long-term reduction of rotavirus incidence under varying levels of coverage. The ensemble of models predicts that the current burden of severe rotavirus disease is 2.6–3.7% of the population each year and that a 2-dose vaccine schedule achieving 70% coverage could reduce burden by 39–42%.

Kernel Machine and Distributed Lag Models for Assessing Windows of Susceptibility to Mixtures of Time-Varying Environmental Exposures in Children's Health Studies

Ander Wilson
Colorado State University
E-mail: ander.wilson@colostate.edu

Abstract: Research has shown that early life exposures to environmental chemicals, starting as early as conception, can reprogram developmental trajectories to result in altered health status later in life. These principles likely apply to complex mixtures as well as individual chemicals. We thus consider statistical methods to estimate the association between mixtures of multiple time-varying exposures and a future health outcome, e.g. exposure to multiple air pollutants observed weekly throughout pregnancy and birth weight. First, we illustrate how to use traditional distributed lag models, distributed lag nonlinear models, and Bayesian kernel machine regression to estimate the association between multiple time-varying exposures and a health outcome. While none of these methods simultaneously accounts for exposure-timing, nonlinear association, and interactions, we highlight situations in which each model performs well. Second, we propose a new method to estimate the association between multiple time-varying exposures and a health outcome. The proposed approach is, to our knowledge, the first method to simultaneously account for exposure-timing, nonlinear associations, and interactions between time-varying exposures. The proposed approach is a Bayesian kernel machine regression method that accounts for exposure timing using a functional weight component within the kernel. The weight function identifies developmental periods with increased association between exposure and a future health outcome, often referred to as a window of susceptibility. We demonstrate the proposed methods in an analysis of exposure to four ambient pollutants and birth weight in a Boston-area perinatal cohort.

S163: New development for statistical analysis

Oracally Efficient Estimation and Simultaneous Inference in Partially Linear Single-index Models for Longitudinal Data

Suojin Wang
Texas A&M University
E-mail: sjwang@stat.tamu.edu

Abstract: Oracally efficient estimation and asymptotically accurate simultaneous confidence band (SCB) are established for the nonparametric link function in the partially linear single-index models for longitudinal data. The proposed procedure works for possibly unbalanced longitudinal data under general conditions. The link function estimator is shown to be oracally efficient in the sense that it is asymptotically equivalent in the order of $n^{-1/2}$ to that with all true values of the parameters being known oracally. Furthermore, the asymptotic distribution of the maximal deviation between the estimator and the true link function is provided, and hence an SCB for the link function is constructed. Finite sample simulation studies are carried out which support our asymptotic theory. The proposed SCB is applied to analyze a CD4 data set.

Modeling and Analysis of Correlated Data using Pairwise Likelihood

Grace Yi
University of Waterloo, University of Western Ontario
E-mail: yyi@uwaterloo.ca

Abstract: Correlated data arise commonly in practice, and modeling and analysis of such data have attracted extensive research interest. Although many methods have been developed, research gaps remain due to emerging issues. For instance, for correlated data with complex structures, modelling complexity is a serious issue, and it is desirable to develop flexible models that are both computationally manageable and interpretatively meaningful. In terms of estimation, much research has been directed to estimation of the mean parameters with the association parameters treated as nuisance. There is relatively less work concerning both the marginal and association

structures, especially in the semiparametric framework. In this talk, I will describe some methods of handling correlated data which are developed based on the pairwise likelihood formulation. Associated modeling strategies and estimation procedures will be discussed.

On the asymptotic distribution of model averaging based on information criterion

Guohua Zou
Capital Normal University
E-mail: ghzou@amss.ac.cn

Abstract: Smoothed AIC (S-AIC) and Smoothed BIC (S-BIC) weight s are widely used in model averaging. In this paper, we investigate t heir asymptotic behavior. The asymptotic distributions of the correspo nding model averaging estimators, the S-AIC and S-BIC estimators, u nder the general fixed parameter framework are derived. Further, we t heoretically check the confidence intervals constructed by Buckland et al. (1997), which have not been studied from theoretical aspect in lit erature. Both simulation study and real data analysis support our theo retical conclusions.

S164: New Advances in Complex Data Analysis and the Applications

Regularization of High-Dimensional Toeplitz Covariance Structure via Entropy Loss Function

Jianxin Pan
Sichuan University
E-mail: Jianxin.Pan@manchester.ac.uk

Abstract: The estimation of structured covariance matrix arises in many applications. An appropriate covariance structure not only improve the accuracy of covariance estimation but also increase the efficiency of the mean parameter estimation in statistical models. For example, a good estimation of covariance structures leads to accurate trajectory predictions for longitudinal data and time series data. In this paper we propose a novel statistical method that is able to select the optimal Toeplitz structure and estimate the high-dimensional covariance matrix simultaneously. Entropy loss functions with nonconvex penalties are employed as matrix-discrepancy measures, under which the optimal covariance structure and the selection of the associated Toeplitz structures are made, simultaneously. The resulting Toeplitz structured covariance estimators are guaranteed to be positive definite, unbiased and selection consistent. Asymptotic theories are derived and simulation studies are conducted, showing a very high accurate Toeplitz covariance structure estimation. The proposed method is also applied to three real data practices, demonstrating its good performance in covariance estimation in practice.

Feature screening with censored data

Qihua Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
E-mail: qhwang@amss.ac.cn

Abstract: Feature screening methods are developed under random censorship.

Conditional Quantile Random Forest with its application for predicting the risk (Post-Traumatic Stress Disorder) PTSD after experienced an acute coronary syndrome

Huichen Zhu
The Hong Kong University of Science and Technology
E-mail: hz2366@cumc.columbia.edu

Abstract: Classification and regression trees (CART) are a classic statistical learning method that efficiently partitions the sample space into mutually exclusive subspaces with the distinctive means of an outcome of interest. It is a powerful tool for efficient subgroup analysis and allows for complex associations and interactions to achieve high prediction accuracy and stability. Hence, they are appealing tools for precision health applications that deal with large amounts of data from EMRs, genomics, and mobile data and aim to provide a transparent decision mechanism. Although there is a vast literature on decision trees and random forests, most algorithms identify subspaces with distinctive outcome means. The most vulnerable or high-risk groups for certain diseases are often patients with extremely high (or low) biomarker and phenotype values. However, means-based partitioning may not be effective for identifying patients with extreme phenotype values. We propose a new regression tree framework based on quantile regression that partitions the sample space and predicts the outcome of interest based on conditional quantiles of the outcome variable. We implemented and evaluated the performance of the conditional quantile trees/forests to predict the risk of developing PTSD after experiencing an acute coronary syndrome (ACS), using an observational cohort data from the REactions to Acute Care and Hospitalization (REACH) study at New York Presbyterian Hospital. The results show that the conditional quantile based trees/forest have better discrimination power to identify patients with severe PTSD symptoms, in comparison to the classical mean based CART. This is joint work with Ying Wei,Katharina Schultebraucks and Ian Kronish.

Large-scale Spatial Predictive Modeling with Applications to Ecological Remote Sensing Data

Chengliang Tang
Columbia University
E-mail: ct2747@columbia.edu

Abstract: Climate change is anticipated to have profound implications for the long-term forest ecosystem resilience and increase carbon fluxes to the atmosphere. Our understanding of these critical problems have been limited by conventional plot-level ground-based observations. Recently developed remote sensing techniques such as stereo photos provide large-scale high-resolution data that would allow researchers to carry out ecological surveys of forest at an unprecedented scale. In this work, we propose a data science workflow to predict palm tree density on forested landscapes in Puerto Rico by applying machine learning tools to data from imaging and remote sensing technologies, combined with ground observation data from field plots. Also, we perform data analysis based on the palm density prediction to reveal its link with other environmental factors, and verify existing ecological theories about landscape characteristics.

S165: New Statistical Challenges in Biomedical Research

Weighted multiple-quantile classifiers for functional data with application in multiple sclerosis screening

Catherine Liu
The Hong Kong Polytechnic University
E-mail: macliu@polyu.edu.hk

Abstract: Multiple sclerosis (MS) is the most prevalent chronic neurological disease. It can be diagnosed by functional data generated from diffusion tensor imaging. Early recognition and treatment of MS are crucial in the treatment and management of MS patients. Existing functional

classifiers seem to suffer from high false negative rates or high false positive rates or both. To develop a classifier with low false negative and false positive rates, we define a generalized distance measure for the functional data. Using this generalized distance, we show that the existing classifiers can be derived by choosing appropriate loss functions. Furthermore, when we consider the quantile loss function, we are able to develop a weighted multiple-quantile (weMulQ) classifier that is robust, accurate, and computationally fast. We showed that it is asymptotically consistent and enjoys the near perfection optimality. Numerically, we demonstrate that it outperforms the other methods when the data are from a generalized Gaussian noise process with mixed populations. Finally, we apply weMulQ to classify MS patients using a DTI data set collected from the Johns Hopkins University and the Kennedy-Krieger Institute. Our classifier indeed has much lower false negative and false positive rates than the existing methods.

Analysis of semi-competing risks data using Archimedean copula models

Antai Wang
New Jersey Institute of Technology
E-mail: antai.wang@njit.edu

Abstract: In this talk, we propose a new method to analyze semi-competing risks data using Archimedean copula models. Our method is simpler than the existing one and tends to be less biased. We illustrate our method using an example.

Two-way partial AUC and its properties

Hanfang Wang
Renmin University of China
E-mail: hyang@ruc.edu.cn

Abstract: Simultaneous control on true positive rate and false positive rate is of significant importance in the performance evaluation of diagnostic tests. Most of the established literature utilizes partial area under receiver operating characteristic (ROC) curve with restrictions only on false positive rate (FPR), called FPR pAUC, as a performance measure. However, its indirect control on true positive rate (TPR) is conceptually and practically misleading. In this paper, a novel and intuitive performance measure, named as two-way pAUC, is proposed, which directly quantifies partial area under ROC curve with explicit restrictions on both TPR and FPR. To estimate two-way pAUC, we devise a nonparametric estimator. Based on the estimator, a bootstrap-assisted testing method for two-way pAUC comparison is established. Moreover, to evaluate possible covariate effects on two-way pAUC, a regression analysis framework is constructed. Asymptotic normalities of the methods are provided. Advantages of the proposed methods are illustrated by simulation and Wisconsin Breast Cancer Data. We encode the methods as a publicly available R package tpAUC.

Tracy-Widom law for the largest eigenvalue of sample covariance matrix generated by VARMA

Yangchun Zhang
Harbin Institute of Technology
E-mail: 540013570@qq.com

Abstract: In this paper we derive the Tracy-Widom law for the largest eigenvalue of sample covariance matrix generated by the vector autoregressive moving average model when the dimension is comparable to the sample size. This result is applied to making inference on the vector autoregressive moving average model. Simulations are conducted to demonstrate the finite sample performance of our inference.

S166: Challenges and Analysis of Complex Data

Classified mixed logistic model prediction

Hanmei Sun
Shandong Normal University
E-mail: hmsun@mail.sdu.edu.cn

Abstract: "We develop a classified mixed logistic model prediction (CMLMP) method for clustered binary data by extending a method proposed by Jiang et al. (2018, J. Amer. Statist. Assoc.) for continuous outcome data. By identifying a class, or cluster, that the new observations belong to, we are able to improve the prediction accuracy of a probabilistic mixed effect associated with a future observation over the traditional method of logistic regression and mixed model prediction without matching the class. Furthermore, we develop a new strategy for identifying the class for the new observations by utilizing covariates information, which improves accuracy of the class identification. In addition, we develop a method of obtaining second-order unbiased estimators of the mean squared prediction errors (MSPEs) for CMLMP, which are used to provide measures of uncertainty. We prove consistency of CMLMP, and demonstrate finite-sample performance of CMLMP via simulation studies. Our results show that the proposed CMLMP method outperforms the traditional methods in terms of predictive performance. An application to medical data is discussed. This work is joint with Thuan Nguyen of Oregon Health & Science University, USA, Yihui Luan of Shandong University, China and Jiming Jiang of University of California, Davis, USA."

Examining causal effects of treatments for patients with infective endocarditis

Yingwei Peng
Queen's University
E-mail: yingwei.peng@queensu.ca

Abstract: A marginal structural Cox's model is employed to explore the causal effect of surgical treatment of infective endocarditis when compared to medical treatment. The investigation is based on the Institute for Clinical Evaluative Sciences (ICES) administrative healthcare databases. Our study is aimed to address controversies among the apparent protective effect of surgery and whether selection bias accounts for the effect.

Latent Class Modeling of Longitudinal Biomarkers in Patients with Chronic Kidney Diseases

Peter Yang
University of Pennsylvania
E-mail: weiyang@pennmedicine.upenn.edu

Abstract: Patients with Chronic Kidney Diseases (CKD) are closely monitored by repeatedly measuring their kidney biomarkers, e.g., estimated glomerular filtration rate (eGFR) and urine protein, over time. The progression of CKD is quite heterogeneous and often corresponds to different etiologies. The identification of different types of CKD progression and associated risk factors are of great interest in CKD patient care. In this project, we proposed latent class mixed effects models to jointly model the longitudinal trajectories of eGFR and urine protein; and identified different CKD progression types. Using Bayes rule, we classified patients in different subgroups and examined class-specific risk factors associated with the progression of CKD.

Health outcomes research with electronic health records: opportunities and challenges

Guofen Yan
University of Virginia
E-mail: guofen.yan@virginia.edu

Abstract: Using electronic health records (EHR) to conduct health outcomes research presents unique opportunities to address research questions that have not been possible with the randomized study design, such as the relationship of race and outcomes. In this presentation, we use chronic kidney disease (CKD) as a case study to demonstrate great opportunities and challenges in the use of EHRs. CKD is a major public health problem in the US and worldwide, with the current estimated prevalence of 10-12% in the US. Many questions remain unanswered including the reasons behind the substantial difference in survival rates among racial and ethnic groups with dialysis treatment. Our project utilizes the U.S. veteran EHR database for >8.9 million veterans, the largest EHR database among U.S. health care systems. We present our study design, some modeling and analysis approaches including an event history analysis perspective, with accompanying challenges such as disease/outcome states ascertained from irregular follow-up visits.

S167: New Advances on Complex Data Analysis

Penalized Empirical Likelihood for the Sparse Cox Model

Yichuan Zhao
Georgia State University
E-mail: yichuan@gsu.edu

Abstract: The current penalized regression methods for selecting predictor variables and estimating the associated regression coefficients in the Cox model are mainly based on partial likelihood. In this paper, an empirical likelihood method is proposed for the Cox model in conjunction with appropriate penalty functions when the dimensionality of data is high. Theoretical properties of the resulting estimator for the large sample are proved. Simulation studies suggest that empirical likelihood works better than partial likelihood in terms of selecting correct predictors without introducing more model errors. The well-known primary biliary cirrhosis data set is used to illustrate the proposed empirical likelihood method. Joint work with Dongliang Wang and Tong Tong Wu.

Retrospective score tests versus prospective score tests for genetic association with case-control data

Yukun Liu
East China Normal University
E-mail: ykliu@sfs.ecnu.edu.cn

Abstract: Since the seminal work by Prentice and Pyke (1979), the prospective logistic likelihood has become the standard method of analysis for retrospectively collected case-control data, in particular for testing the association between a single genetic marker and a disease outcome in genetic case-control studies. When studying multiple genetic markers with relatively small effects, especially those with rare variants, various aggregated approaches based on the same prospective likelihood have been developed to integrate subtle association evidence among all considered markers. In this paper we show that using the score statistic derived from a prospective likelihood is not optimal in the analysis of retrospectively sampled genetic data.We develop the locally most powerful genetic aggregation test derived through the retrospective likelihood under a random effect model assumption. In contrast to the fact that the disease prevalence information cannot be used to improve the efficiency for the estimation of odds ratio parameters in logistic regression models, we show that it can be utilized to enhance the testing power in genetic association studies. Extensive simulations demonstrate the advantages of the proposed method over the existing ones. One real genome-wide association study is analyzed for illustration.

Brain-wide organizations of neuronal activity in larval zebrafish

Yu Hu
The Hong Kong University of Science and Technology
E-mail: mahy@ust.hk

Abstract: Simultaneous recordings of large populations of neurons in behaving animals allow detailed observation of high-dimensional, complex brain activity. Here we focus on a dataset where calcium activity signals of individual neurons are recorded near-simultaneously across the whole-brain in larval zebrafish. At the same time, multiple visual stimuli are presented to the animal to induces a variety of behaviors. By comparing neurons' activity with stimuli and behaviors, we identified a group of neurons coding for multiple stimulus features that elicit similar behavioral responses. To study brain-wide activity beyond explicit sensorimotor processing, we used an unsupervised clustering technique that organizes neurons into groups with similar activity. The analysis recovers known brain nuclei as well as complexity of functional clusters in terms of unclustered neurons, heterogeneity, and anatomical structures. Recent works on towards identifying stable clusters across multiple conditions will also be discussed.

Interaction Pursuit Biconvex Optimization

Yuehan Yang
Central University of Finance and Economics
E-mail: yyh@cufe.edu.cn

Abstract: We study the high-dimensional multivariate regression analysis with the number of predictors and the number of responses both grow at an exponential rate in sample size. The proposed method explores the regression relationship when the predictors, errors and responses are all assumed be the samples of different multivariate normal distributions with general covariance matrices. We use the precision matrix estimation for multivariate analysis and use the laplacian quadratic associated with the graph information to promote smoothness among coefficients associated with the correlated predictors and responses. Theoretical results are proved under interpretable conditions. We provide an efficient algorithm for computing the estimates. Simulation studies and real data examples compare the proposed methods with several existing methods, indicating that the proposed methods achieve better interpretability and accuracy.

S168: Complex Data Analysis in Business, Economics and Industry

Quality Big Data

Fugee TSUNG
HKUST
E-mail: season@ust.hk

Abstract: This talk will present and discuss the challenges and opportunities that quality engineers and managers face in the era of big data. The ability to separate signal and noise in the data-rich-information-poor environment would be the key, especially for industrial big data. Emerging research issues include data fusing with heterogeneous data sources, statistical transfer learning, and statistical process control and monitoring

for big data streams.

On selecting valid instruments for structural vector autoregression

CY (Chor-yiu) SIN
National Tsing Hua University
E-mail: cysin@mx.nthu.edu.tw

Abstract: With the prevalence of the so-called "big data", structural models/equations are often estimated with high-dimensional instruments. Notable research papers include Belloni, Chen, Chernozhukov and Hansen (2012); and Kang, Zhang, Cai and Small (2016). The former assumes all instruments are valid and considers an efficient estimator; while the latter proposes some confidence sets of the structural parameters, and investigates their properties under various assumptions on the number of valid instruments. In this paper, we adopt and modify the OGA-HDIC algorithm proposed by Ing (2019) and search for valid instruments out of some high-dimensional potential instruments. Unlike Lasso, this algorithm is arguably more suitable for time-series data. We close this paper with (i) Some comparisons with the high-dimensional Durbin-Wu-Hausman (DWH) test proposed by Guo, Kang, Cai and Small (2018); (ii) Some Monte-Carlo simulations.

Mean squared prediction errors of integrated autoregressive models with polynomial time trends

Shu-Hui Yu
National University of Kaohsiung
E-mail: shuhui@nuk.edu.tw

Abstract: Assume that observations are generated from nonstationary autoregressive (AR) processes with deterministic time trends. We adopt a fitted model which possibly over specifies the orders of autoregressive and time trends to predict future observations and obtain an asymptotic expression for the multistep mean-squared prediction error (multistep MSPE) of the least squares predictor. This expression provides the first exact assessment of the impacts of nonstationarity, model complexity, and model over-specification on the corresponding multistep MSPE. It not only provides a deeper understanding of the least squares predictors in nonstationary time series, but also forms the theoretical foundation for asymptotical efficient order selection in nonstationary AR processes with possibly deterministic time trends.

Hing-dimensional model selection under covariate shift

Ching-Kang Ing
National Tsing Hua University
E-mail: cking@stat.nthu.edu.tw

Abstract: We consider a high-dimensional variable selection problem under covariate shift, which has been paid much attention in statistics and machine learning fields. We first use the orthogonal greedy algorithm (OGA) to implement model selection and then adjust the selected model for covariate shift using an importance weighted OGA (IWOGA). Under a weak sparsity condition on regression coefficients, a rate of convergence of IWOGA is derived. In addition, the performance of IWOGA under the covariate shift is demonstrated through a simulation study and a real data example.

S169: Statistical Methods and Theory for Complex and Large Data

How to apply the multilevel modeling in large health care administrative data

Jun Guan
Methodologist
E-mail: jun.guan@ices.on.ca

Abstract: In health care research, we often need to analyze the hierarchical or multi-level, or nested data. For example, long-term care residents living in the same nursing home have access to similar physicians and nurses with similar prescribing practices. Patients operated on by the same surgeon have outcomes that are more similar than those operated on by a different surgeon since outcome partly depends on the surgeon's skill; furthermore, patients of surgeons working at the same hospitals may have similar outcomes because post-operative care varies among hospitals. In this talk, I will share some experience on how the multilevel modeling has been applied in large health administrative data, plan to address some challenges when we are analyzing the health care big data (millions records), and demonstrate some tips/tricks on how we have handled multilevel modeling in some software such as SAS, R, STATA.

Keywords: multilevel modeling, health administrative data, SAS, R, STATA.

Panel Data Models with Potentially Misspecified Unknown Factors

Huanjun Zhu
Xiamen University
E-mail: hzhu928@xmu.edu.cn

Abstract: While studying panel data models with interactive fixed effects, majority works focus on generalizing how regressors and the associated marginal effects enter the models (e.g., non-/semi- parametric form, heterogeneity setting, etc). However, few effort has been made to generalise the unobservable factor structure and relax the corresponding assumptions. In this study, we investigate the consequences of misspecifying the property of unknown factors of a parametric panel data model. We show that the interactive fixed effects estimator still achieves the global minimum even when the properties of unknown factors are misspecified. Some rates of convergence and an asymptotic normality are established accordingly. In addition, we find that nonstationarity of the factors can help reduce the requirement of the sample size along the time dimension. Moreover, the investigation on misspecification extends the discussions of Section 4.2 of Bai et al. (2009). Finally, we verify our findings through extensive simulation studies, and investigate income elasticity of health care expenditure using data of OECD counties.

Dynamic Functional Connectivity Change-point Detectionbased on Random Matrix Theory

Jaehee Kim
Duksung Women's University
E-mail: jaehee@duksung.ac.kr

Abstract: Most statistical analyses of fMRI data assume that the nature, timing and duration of the psychological processes in the controlled experiment. However, it is often hard to specify this information a priori. In this work we introduce develop a method that leverages the special structure of our covariance model with regions of interest (ROIs). The technique enables relatively fast and efficient change-point estimation.

Estimating the eigenvalues of a population covariance matrix from a sample covariance matrix is a problem of fundamental importance in multivariate statistics; the eigenvalues of covariance matrices a key role in many widely used techniques. We apply the Tracy-Widom transformation

(Tracy and Widom 1996) for the largest eigenvalue of the covariance matrix ratio up to time t, and that after time t, for each time-point t. For resting state fMRI data, the covariance function is probably one of the most important quantities of interest. Change-point analysis based on the maximum eigenvalue and canonical correlation approach are useful tools in situations where high-dimensional data are collected. We further examine dynamic FC properties by estimating change-points and performing group comparisons. Using our proposed method, we conduct simulation study and analyze fMRI data from a study of epilepsy patients. The method is applied to various simulated data sets as well as to an fMRI data set from epilepsy study.

Index of Authors

Index of Authors

Index of Authors

Index of Authors

See you at the 2020 ICSA China Conference, Wuhan, China

www.icsa.org

International Chinese Statistical Association

泛華統計協會