2025 ICSA China Conference

Zhuhai, Guangdong, China

June 28 - June 30, 2025

International Chinese Statistical Association

泛華統計協會

International Chinese Statistical Association

China Conference
2025

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 28 - June 30, 2025
Beijing Normal University
Zhuhai, Guangdong, China

Organized by

International Chinese Statistical Association

Hosted by

Faculty of Arts and Sciences, Beijing Normal University
School of Statistics, Beijing Normal University

©2025

International Chinese Statistical Association

Content

ICSA 2025 China Conference

June 28- June 30, Zhuhai, Guangdong, China

Welcome to the International Chinese Statistical Association (ICSA) 2025 China Conference!

The ICSA 2025 China Conference will be held in Zhuhai, Guangdong, China, from June 28 to June 30, 2025. It is co-organized by ICSA and Beijing Normal University. The conference venue is located at Beijing Normal University, Zhuhai Campus (BNUZ). The theme of this conference is the ***"Future in Statistics: Partnership and Innovation in the Data-Rich Era",*** aiming to explore cutting-edge development, interdisciplinary collaboration, and innovative applications of statistics and data science in the era of big data.

The executive and organization committees have been working diligently to put together a robust and comprehensive program, including two keynote lectures, 112 invited sessions, eight contributed sessions, a junior researcher award winners session, 17 poster presentations, and exciting social events. Keynote lectures are from leading experts in statistics, ***Dr. Runze Li*** (Pennsylvania State University, USA) and ***Dr. Wenguang Sun*** (Zhejiang University, China).

With your full support, this conference attracts more than 550 statisticians and data scientists working in academia, government, and industry from all over the world. We hope that the conference provides great opportunities for learning, networking, collaborations and recruiting. You will share the thoughts and ideas with conference guests, and receive inspirations from old research ideas and develop new ones. The local organization committee and more than 50 student volunteers led by Professor Shumei Zhang in BNUZ have made a great effort to arrange the meeting logistics and social events in this conference, including the banquet (Sunday, June 29, evening, banquet speaker ***Dr. Tian Zheng***). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

The city of Zhuhai enjoys a subtropical climate, offering year-round comfort, and is easily accessible via high-speed rail, ferries, and international airports. This coastal gem offers pristine beaches, luxury resorts, and vibrant shopping districts. Famed for the Hong Kong-Zhuhai-Macao Bridge (the world’s longest sea crossing) and Chimelong Ocean Kingdom, Zhuhai blends natural beauty with modern thrills. Its Cantonese seafood cuisine is a must-try. As a seaside garden city, Zhuhai invites you to unwind or explore, promising an unforgettable escape between ocean vistas and urban charm.

Thanks for coming to the ICSA 2025 China Conference in Zhuhai!

Yuanjia Wang, on behalf of
ICSA 2025 China Conference Executive and Organization Committees

EXECUTIVES

President: Hongyu Zhao (hongyu.zhao@yale.edu)

Past President: Xun Chen (xun.chen@abbvie.com)

President-Elect: Rong Chen (rongchen@stat.rutgers.edu)

Executive Director: Jun Zhao (2023-2025, executive.director@icsa.org)

ICSA Treasurer: Xin He (2025-2027, treasurer@icsa.org)

The ICSA Office Manager: Grace Ying Li

Email: oicsa@icsa.org, Phone: (317) 287-4261

BOARD OF DIRECTORS

Ming Tan (2023-2025, mtt34@georgetown.edu)

Huazhen Lin (2023-2025, linhz@swufe.edu.cn)

Min Zhang (2023-2025, mzhangst@umich.edu)

Li Wang (2023-2025, li.wang1@abbvie.com)

Yanping Wang (2023-2025, WANG_YANPING@LILLY.COM)

Jialiang Li (2024-2026, jialiang@nus.edu.sg)

George Tseng (2024-2026, ctseng@pitt.edu)

Kun Chen (2024-2026, kun.chen@uconn.edu)

Song Yang (2024-2026, yangso@nhlbi.nih.gov)

Jianchang Lin (2024-2026, Jianchang.Lin@takeda.com)

Yuguo Chen (2025-2027, yuguo@illinois.edu)

Haitao Chu (2025-2027, hchu.xy@gmail.com)

Yuanjia Wang (2025-2027, yw2016@cumc.columbia.edu)

Xinping Cui (2025-2027, xpcui@ucr.edu)

Yushi Liu (2025-2027, liu_yushi@lilly.com)

COMMITTEES

Program Committee

Chair: Qingxia Chen (2025, cindy.chen@vanderbilt.edu)

Xinping Cui (2023-2025, ICSA China Conference 2024, ICSA International Conference 2025, xpcui@ucr.edu)

Yuanjia Wang (2024-2026, ICSA China Conference 2025, yw2016@cumc.columbia.edu)

Ying Zhang (2024-2026, ICSA International Conference 2025, ying.zhang@unmc.edu)

Ming-Chung Chang (2024-2026, ICSA International Conference 2025, mcchang@stat.sinica.edu.tw)

Dandan Liu (2023-2025, ICSA Symposium 2024, dandan.liu@vumc.org)

Dacheng Liu (2024-2026, ICSA Symposium 2025, dacheng.liu@boehringer-ingelheim.com),

Xiaojing Wang (2024-2026, ICSA Symposium 2025, xiaojing.wang@uconn.edu)

Lily Wang (2025-2027, ICSA Symposium 2026, lwang41@gmu.edu)

Wanli Qiao (2025-2027, ICSA Symposium 2026, wqiao@gmu.edu)

Yingwen Dong (2023-2025, JSM Representative 2024, yingwen.dong@sanofi.com)

Shuangge Ma (2024-2026, JSM Representative 2025, shuangge.ma@yale.edu)

Awards Committee

Chair: Yong Chen (2025, ychen123@pennmedicine.upenn.edu)

Zhigang Li (2023-2025, zhigang.li@ufl.edu)

Chunming Zhang (2023-2025, cmzhang@stat.wisc.edu)

Wei Wu (2023-2025, wwu@fsu.edu)

Lu Tian (2023-2025, lutian@stanford.edu)

Xuezhou Mao (2023-2025, Xuezhou.Mao@modernatx.com)

Hong Tian (2024-2026, hong.tian@beigene.com)

Huilin Li (2024-2026, Huilin.Li@nyulangone.org)

Gen Li (2024-2026, ligen@umich.edu)

Kai Yang (2024-2026, kayang@mcw.edu)

Nominating and Election Committee

Chair: Hongjian Zhu (2025, Hongjian.zhu@abbvie.com)

Yichuan Zhao (2023-2025, yichuan@gsu.edu)

Wenqing He (2023-2025, whe@stats.uwo.ca)

Zhigang Li (2023-2025, zhigang.li@ufl.edu)

Tiejun Tong (2024-2026, tongt@hkbu.edu.hk)

Li-Shan Huang (2024-2026, lhuang@stat.nthu.edu.tw)

Wei Zhang (2024-2026, wei.zhang@boehringer-ingelheim.com)

Jin-Ting Zhang (2024-2026, stazjt2020@nus.edu.sg)

Jingshen Wang (2025-2027, jingshenwang@berkeley.edu)

Min Chen (2025-2027, mchen@utdallas.edu)

Jimin Ding (2025-2027, jmding@wustl.edu)

Special Lecture Committee

Chair: Hongzhe Lee (2025, hongzhe@pennmedicine.upenn.edu)

Ming Tony Tan (2023-2024, mtt34@georgetown.edu)

Ronghui Xu (2025-2027, rxu@health.ucsd.edu)

Zijian Guo (2025-2027, rxu@health.ucsd.edu)

Jianxin Shi (2025-2027, jianxin.shi.nci@gmail.com)

Bingming Yi (2025-2027, Bingming_Yi@vrtx.com)

Publication Committee

Chair: Runze Li (2025, rzli@psu.edu)

Lan Wang (2025-2027, wangx346@gmail.com)

Zhigen Zhao (2025-2027, zhigen.zhao@temple.edu)

Meg Gamalo (2025-2027, Co-Editors of SIB, Margaret.Gamalo@pfizer.com)

Jianguo (Tony) Sun (2024-2026, Co-Editors of SIB, sunj@missouri.edu)

Yi-Hau Chen (2023-2026, Co-Editors of Statistica Sinica, yhchen@stat.sinica.edu.tw)

John Stufken (2023-2026, Co-Editors of Statistica Sinica, jstufken@gmu.edu)

Huixia Judy Wang (2023-2026, Co-Editors of Statistica Sinica, judywang@gwu.edu)

Ding-Geng (Din) Chen (Co-Editor of ICSA book series, dinchen@email.unc.edu)

Yichuan Zhao (2025-2027, Co-Editor of ICSA book series, yichuan@gsu.edu)

Chixiang Chen (2024-2026, Editor for ICSA Bulletin, chixiang.chen@som.umaryland.edu)

Jun Zhao (2023-2025, Executive Director of ICSA, executive.director@icsa.org)

Grace Ying Li (2024-2026, Editor of ICSA Newsletter, li_ying_grace@lilly.com)

Membership Committee

Chair: Tiejun Tong (2025, tongt@hkbu.edu.hk)

Victoria Chang (2025-2027, victoria.chang@beigene.com)

Pengfei Li (2025-2027, pengfei.li@uwaterloo.ca)

Tu Xu (2023-2025, xutu1116@gmail.com)

Zhigen Zhao (2025-2027, zhaozhg@temple.edu)

Wangli Xu (2025-2027, wlxu@ruc.edu.cn)

Yunwen Yang (2025-2027, andrea.yang2@gmail.com)

IT Committee

Chair: Chengsheng Jiang (2025, website@icsa.org)

Archive Committee

Chair: Jun Yan (2025, jun.yan@uconn.edu)

Raymond Wong (2025-2027, raywong@tamu.edu)

Finance Committee

Chair: Xin He (2025-2027, xinhe@umd.edu)

Rochelle Fu (2022-2025, fur@ohsu.edu)

Rui Feng (2022-2025, ruifeng@pennmedicine.upenn.edu)

Financial Advisory Committee

Chair: Xiangqin Cui (2025, xiangqin.cui@emory.edu)

Fang Chen (2022-2025, FangK.Chen@sas.com)

Nianjun Liu (2020-2025, liunian@indiana.edu)

Rochelle Fu (2020-2025, fur@ohsu.edu)

Yuan Jiang (2020-2025, yuan.jiang@stat.oregonstate.edu)

Hongliang Shi (2020-2025, hongliangshi15@gmail.com)

ICSA Representative to JSM Program Committee

Yingwen Dong, (2024, yingwen.dong@sanofi.com)

Shuangge Ma (2025, shuangge.ma@yale.edu)

ICSA Outreach and Engagement Committee

Co-Chairs: Jin Zhou (2023-2025, jinjinzhou@g.ucla.edu)

Qing Yang (2023-2025, qing.yang@duke.edu)

Weining Shen (2023-2025, weinings@uci.edu)

Xiaowu Dai (2025-2027, dai@stat.ucla.edu)

Yuhua Zhu (2025-2027, yuhuazhu@ucla.edu)

Yu Wu (2025-2027, yuwu58@gmail.com)

Hua Zhou (2025-2027, huazhou@ucla.edu)

ICSA Constitution Committee

Chair: Hongzhe Lee (2023-2025, hongzhe@pennmedicine.upenn.edu)

Rochelle Fu (2023-2025, fur@ohsu.edu)

Jun Zhao (2023-2025, executive.director@icsa.org)

Ying Lu (2023-2025, ylu1@stanford.edu)

Jianguo Sun (2023-2025, sunj@missouri.edu)

Chengsheng Jiang (2023-2025, website@icsa.org)

Yichuan Zhao (2023-2025, yichuan@gsu.edu)

Jiayang Sun (2023-2025, jsun21@gmu.edu)

Webinar Sub-Committee

Co-Chairs: Qing Yang (2025, qing.yang@duke.edu)

Deli Wang (2025, wangdeli@gmail.com)

Jun Zhao (2023-2025, executive.director@icsa.org)

Helena Fan (2023-2025, Helena.fan@tlgcareers.com)

Lei Wang (2023-2025, Lei.Wang@tlgcareers.com)

Xiyuan Gao (2024-2026, sheila.gao@lilly.com)

Jessica (Jingxian) Cai (2024-2026, jingxian.cai@regeneron.com)

Meng Wang (2024-2026, wm.mengwang11@gmail.com)

Betty (Yutong) Tan (2024-2026, bettytan9614@gmail.com)

CONFERENCE COMMITTEES

2025 Applied Statistics Symposium

Co-Chairs: Xiaojing Wang (xiaojing.wang@uconn.edu)
Dacheng Liu (dacheng.liu@boehringer-ingelheim.com)

2025 ICSA China Conference

Co-Chairs: Yuanjia Wang (yw2016@cumc.columbia.edu)
Lixing Zhu (lzhu@bnu.edu.cn)

2025 JSM Local Committee

Chair: Dandan Liu (dandan.liu@vumc.org)

2025 ICSA International Conference

Co-Chairs: Ying Zhang (ying.zhang@unmc.edu)
Xinping Cui (xpcui@ucr.edu)
Ming-Chung Chang (mcchang@stat.sinica.edu.tw)

2026 Applied Statistics Symposium

Co-Chairs: Lily Wang (lwang41@gmu.edu),
Wanli Qiao (wqiao@gmu.edu)

ICSA CHAPTERS

ICSA-Taiwan Chapter

Henry Horng-Shing Lu (Chair, hslu@stat.nycu.edu.tw)

Chao A. Hsiung (Past Chair, hsiung@nhri.org.tw)

ICSA-Canada Chapter

Wenqing He (Chair, whe@stats.uwo.ca)

Joan Hu (Past Chair, joan_hu@sfu.ca)

Leilei Zeng (Secretary/Treasurer, lzeng@uwaterloo.ca)

ICSA-Midwest Chapter

Ziqian Geng (Chair, ziqian.geng@abbvie.com)

Xiaohong Huang (Past Chair, xiaohong.huang@abbvie.com)

Conference Committees

Executive Committee

Co-Chairs: Lixing Zhu, Beijing Normal University

Co-Chairs: Yuanjia Wang, Columbia University

Xun Chen, ICSA past president

Xingwei Tong, Beijing Normal University

Shumei Zhang, Beijing Normal University

Hongyu Zhao, Yale University

Jun Zhao, Antengene

Scientific Program Committee

Co-Chairs: Lixing Zhu, Beijing Normal University, Committee Co-Chair

Co-Chairs: Yuanjia Wang, Columbia University, Committee Co-Chair

Mingyao Ai, Peking University

Fei Chen, Yunnan University of Finance and Economics

Mingyen Cheng, Hong Kong Baptist University

Xinping Cui, University of California, Riverside

Yifan Cui, Zhejiang University

Yixin Fang, AbbVie

Xingche Guo, University of Connecticut

Xu Guo, Beijing Normal University

Zhezhen Jin, Columbia University

Linglong Kong, University of Alberta

Chenlei Leng, University of Warwick

Lexin Li, University of California, Berkeley

Jialiang Li, National University of Singapore

Hua Liang, George Washington University

Huazhen Lin, Southwestern University of Finance and Economics

Lu Lin, Shandong University

Qian Lin, Tsinghua University

Molei Liu, Columbia University

Yanyuan Ma, Penn State University

Guangming Pan, Nanyang Technological University

Jianxin Pan, Beijing Normal University-Hong Kong Baptist University United International College

Chengchun Shi, London School of Economics

Peter Song, University of Michigan, Ann Arbor

Jianguo Sun, University of Missouri

Junhui Wang, The Chinese University of Hong Kong

Tao Wang, Shanghai Jiao Tong University

Zhaojun Wang, Nankai University

Zheyu Wang, Johns Hopkins University

Ruiyang Wu, CUNY Baruch

Shanghong Xie, University of South Carolina

Jinfeng Xu, City University of Hong Kong

Jianfeng Yao, The Chinese University of Hong Kong, Shenzhen

Grace Yi, Western Ontario University

Zhou Yu, East China Normal University

Donglin Zeng, University of Michigan, Ann Arbor

Xin Zhang, Florida State University

Yichuan Zhao, Georgia State University

Shurong Zheng, Northeast Normal University

Wei Zhong, Xiamen University

Liping Zhu, Renmin University of China

Jingjing Zou, University of California, San Diego

Junior Researcher Award Committee

Jinyuan Chang, Southwestern University of Finance and Economics

Qingxia Chen, Vanderbilt University

Yuan Chen, Memorial Sloan Kettering Cancer Center

Yukun Liu, East China Normal University

Zhenke Wu, University of Michigan Ann Arbor

Yuhong Yang, Tsinghua University

IT and Website Committee

Chair: Shumei Zhang, Beijing Normal University, China

Shuangshuang Hou, Beijing Normal University, China

Jiakun Jiang, Beijing Normal University, China

Gaorong Li, Beijing Normal University, China

Qun Liu, Beijing Normal University, China

Local Organization Committee

Co-Chair: Shumei Zhang, Beijing Normal University, China

Co-Chair: Xingwei Tong, Beijing Normal University, China

Feifei Chen, Beijing Normal University, China

Xu Guo, Beijing Normal University, China

Shuangshuang Hou, Beijing Normal University, China

Jiakun Jiang, Beijing Normal University, China

Qing Jiang, Beijing Normal University, China

Yu Jin, Beijing Normal University, China

Gaorong Li, Beijing Normal University, China

Guanxun Li, Beijing Normal University, China

Ran Liu, Beijing Normal University, China

Lei Peng, Beijing Normal University, China

Tao Qiu, Beijing Normal University, China

Zhuo Ren, Beijing Normal University, China

Ning Wang, Beijing Normal University, China

Xiaoyi Wang, Beijing Normal University, China

Chuanlong Xie, Beijing Normal University, China

Lili Xu, Beijing Normal University, China

Zheng Zhai, Beijing Normal University, China

Xiaochen Zhang, Beijing Normal University, China

Huiyan Zhao, Beijing Normal University, China

Junlong Zhao, Beijing Normal University, China

Niwen Zhou, Beijing Normal University, China

Yingxing Li, Xiamen University, China

Jiarui Chi, Sanofi, China

Poster Session Committee

Chair: Xu Guo, Beijing Normal University, China

Yingxing Li, Xiamen University, China

Gaorong Li, Beijing Normal University, China

Chuanlong Xie, Beijing Normal University, China

Ning Wang, Beijing Normal University, China

Jiarui Chi, Sanofi, China

Host Institution & Academic Units

Beijing Normal University

Beijing Normal University (BNU) grew out of the Education Department of Imperial University of Peking established in 1902, which initiated teacher training in China's higher education. After the development for over a century, BNU has become a comprehensive and research-intensive university with its main characteristics of basic disciplines in sciences and humanities, teacher education and educational science.

BNU consists of Beijing Campus and Zhuhai Campus. The University has 3 faculties, 29 schools, 8 research institutes and 5 academies. In addition, there are more than 5.8 million books and 10 million e-books in its libraries. BNU is home to more than 32,000 full-time students, and has 2637 full-time teachers, including 2061 teachers with Senior Professional and Technical Positions.

At present, the university has established cooperative ties with nearly 300 universities and international organizations from over 40 countries and regions. BNU has around 2100 long-term international students, the scale of which ranks among top in China's universities.

(Relevant data as of March, 2025)

北京师范大学

北京师范大学是教育部直属重点大学，是一所以教师教育、教育科学和文理基础学科为主要特色的著名学府。学校的前身是 1902 年创立的京师大学堂师范馆，1908 年改称京师优级师范学堂，独立设校，1912 年改名为北京高等师范学校。1923 年学校更名为北京师范大学，成为中国历史上第一所师范大学。1931 年、1952 年北平女子师范大学、辅仁大学先后并入北京师范大学。

百余年来，北京师范大学始终同中华民族争取独立、自由、民主、富强的进步事业同呼吸、共命运，在“五四”“一二·九”等爱国运动中发挥了重要作用。以李大钊、鲁迅、梁启超、钱玄同、吴承仕、黎锦熙、陈垣、范文澜、侯外庐、白寿彝、钟敬文、启功、胡先骕、汪堃仁、周廷儒等为代表，一大批名师先贤在这里弘文励教。经过百余年的发展，学校秉承“爱国进步、诚信质朴、求真创新、为人师表”的优良传统和“学为人师、行为世范”的校训精神，形成了“治学修身，兼济天下”的育人理念。

“七五”“八五”期间，北京师范大学被确定为国家首批重点建设的十所大学之一。“九五”期间，被首批列入“211 工程”建设计划。2002 年百年校庆之际，教育部和北京市决定重点共建北京师范大学，北京市第九次党代会将北京师范大学列入支持建设的世界一流大学的行列。“十五”期间，学校进入国家“985 工程”建设计划。2017 年，学校进入国家“世界一流大学”建设 A 类名单，11 个学科进入国家“世界一流学科”建设名单。2022 年，学校 12 个学科入选第二轮“双一流”建设学科，入选学科数量位居全国高校前列。

北京师范大学由北京校区、珠海校区两个校区（含五个校园）组成。北京校区现有全日制本科生 9541 人，全日制研究生 12680 人，非全日制研究生 4405 人；珠海校区于 2019 年 4 月由教育部正式批准建设，现有全日制本科生 7002 人，全日制研究生 3616 人，非全日制研究生 1474 人。学校设 3 个学部、29 个学院、8 个研究院（中心）、5 个书院。馆藏印本文献 584 万余册，电子图书 1003 万余册。

北京师范大学学科综合实力位居全国高校前列。2002 年成为首批拥有自主设置本科专业审批权的 6 所高校之一，2018 年成为首批可开展学位授权自主审核的 20 所高校之一。现有本科专业 77 个、一级学科硕士学位授权点 36 个、一级学科博士学位授权点 34 个、专业学位博士授权点 6 个、专业学位硕士授权点 25 个，博士后科研流动站 30 个。

北京师范大学是国家人文社科科研和科技创新的一支重要力量。学校拥有国家高端智库试点单位 1 个、国家重点实验室 4 个、国家工程研究中心 1 个、国家野外科学观测研究站 1 个、国家级协同创新中心 1 个、国家国际科技合作基地 1 个，铸牢中华民族共同体意识研究培育基地 1 个、国家新闻出版署重点实验室 1 个、国家革命文物协同研究中心 1 个、国家语言文字推广基地 1 个，教育部重点实验室 10 个、教育部工程研究中心 7 个、教育部野外科学观测研究站 2 个、教育部国际联合实验室 1 个、教育部人文社会科学重点研究基地 7 个、教育部哲学社会科学实验室 1 个，教育部基础教育质量监测中心 1 个、教育部国别和区域研究基地 4 个、教育部教育立法研究基地 1 个、教育部师德师风建设基地 1 个、教育部战略研究基地 1 个、高等学校学科创新引智基地 9 个，北京高等学校高精尖创新中心 1 个、北京市重点实验室 12 个、北京

市工程技术研究中心 4 个、广东省粤港澳联合实验室 1 个、广东省重点实验室 1 个、广东省野外科学观测研究站 1 个。北京市哲学社会科学重点研究基地 2 个、首都新型高端智库 1 个、北京市语言文字工作委员会研究基地 1 个、北京高校协同创新中心 1 个、北京市习近平新时代中国特色社会主义思想研究中心基地 1 个、北京教育法治研究基地 1 个、北京市中小学师德师风建设基地 1 个，教育部师德师风建设基地 1 个，其他省部级重要平台 26 个。定期出版专业刊物 34 种。

北京师范大学教育资源丰富，是国家高素质创新型人才培养的重要基地。拥有国家文科基础学科人才培养和科学研究基地 2 个、国家理科基础科学研究和教学人才培养基地 5 个、国家基础学科拔尖学生培养计划 2.0 基地 10 个、国家教育体制改革试点学院 1 个；现有国家级实验教学示范中心 4 个、国家级虚拟仿真实验教学中心 2 个；现有国家教材建设重点研究基地 5 个；是国家大学生文化素质教育基地、国家对外汉语教学基地、国家生命科学与技术人才培养基地、国家卓越法律人才教育培养基地、教育部中华优秀传统文化传承基地（中国话剧）；入选国家级一流本科专业建设点 46 个、北京市级一流本科专业建设点 6 个、北京高校“重点建设一流专业”2 个；入选教育部虚拟教研室建设试点 9 个、北京高校虚拟教研室建设试点 3 个；拥有北京市实验教学示范中心 2 个、广东省实验教学示范中心 3 个、广东省基础学科拔尖人才培养创新实验区 1 个。

北京师范大学教师队伍结构合理、素质精良。现有专任教师 2637 人，其中具有高级专业技术职务 2061 人。两院院士 7 人，入选各类国家级重大人才工程 393 人次。

北京师范大学积极服务国家对外开放战略，国际交流合作广泛。2020 年，学校发布《全球发展战略规划》（2020-2025），确立了助力建设全球卓越学术共同体、教育创新共同体、青年发展共同体、高校社会责任共同体的战略愿景。与近 50 个国家和地区的 200 余所大学、研究机构建立了校级合作关系。持续推进与牛津大学、斯坦福大学、赫尔辛基大学等世界一流大学和一流学科的合作，积极拓展与“一带一路”沿线国家的交流，打造了“看中国·外国青年影像计划”等一批具有国际影响力的文化项目，为海内外专家学者和学生开展学术交流、学习深造、文化交流互鉴提供广阔平台。

北京师范大学第十四次党代会提出，要弘扬红色师范百廿传统，坚守教师教育核心使命，不断提升“综合性、研究型、教师教育领先的中国特色世界一流大学”办学水平，以学校的高质量发展全面服务教育强国建设，以教育的高质量发展全面支撑中国式现代化。“十四五”期间，学校正着力构建“高原支撑、高峰引领”的学科发展体系和以北京校区和珠海校区为两翼的一体化办学格局，不断深化综合改革，推进各项事业发展。北京师范大学正向着建设世界一流大学的目标稳步迈进。

Department of Statistics, Factulty of Arts and Sciences

The Faculty of Arts and Sciences at Beijing Normal University is an academic and research institution located on the Zhuhai campus. The Department of Statistics, established in July 2022, is one of the Faculty’s academic divisions.

Since its establishment, the Department of Statistics has achieved notable progress, building a comprehensive talent-training system that covers undergraduate, master's, doctoral, and postdoctoral levels. It has been approved as a Key Construction Discipline in Statistics and Data Science under Guangdong Province’s "Strengthening, Supplementing, and Elevating" Plan (one of the first selected disciplines at the Zhuhai campus), as well as the Guangdong Provincial Key Laboratory of Educational Psychology and Data Science Technology & Applications. Additionally, the Department has led the founding of the Guangdong–Hong Kong–Macao Alliance of Universities for Statistics and Data Science.

Currently, the Department is dedicated to addressing international frontier issues in statistics and data science and meeting significant strategic demands of the data technology industry in the Guangdong–Hong Kong–Macao region. Focusing on distinctive disciplinary directions such as foundational theories of statistics and computational technology, management and evaluation methods for educational big data, and modeling and analysis of complex structured data, the Department aims to build a statistics and data science discipline with extensive international influence, distinct from the School of Statistics at the Beijing campus while being firmly rooted in the Zhuhai campus.

The Department currently employs 19 full-time faculty members, all holding doctorates: 5 Professors, 7 Associate Professors/Associate Research Fellows, and 7 Lecturers. Six faculty members have earned their doctorates abroad and all junior faculty members have overseas study experience. All faculty members excel in both teaching and research. Professor Lixing Zhu, the academic leader, was the first recipient in statistics of the National Science Fund for Distinguished Young Scholars, is a Changjiang Chair Professor, serves on the Ministry of Education’s Teaching Steering Committee for Statistics, and is a Fellow of the AAAS, the American Statistical Association, and the Institute of Mathematical Statistics, as well as an Elected Member of the International Statistical Institute. Moreover, he has independently won several prestigious domestic and international awards, including the State Natural Science Award (Second Class) and the Humboldt Research Award from Germany. Professors Yong Li and Shumei Zhang have received the Beijing Higher Education Teaching Master Award, the Baosteel Excellent Teacher Award, the Golden Medal for “Four Good Teachers” at Beijing Normal University, and Second Prize in the Beijing Higher Education Teaching Achievement Awards. Associate Professor Chuanlong Xie was selected in 2024 for Guangdong’s TZ Plan (Science and Technology Platforms) as an Outstanding Young Talent. Young faculty members regularly publish in leading journals such as Nature Communications, JASA, Biometrika, IEEE TPAMI, JMLR, and the JoE and have won first and second prizes in the University's Young Teachers Competition as well as the Peng Nian Distinguished Young Teacher Award. Since 2020, the Department has secured one Key Project of the National Natural Science Foundation of China (the first on the Zhuhai campus), two General Projects and ten Young Scientist Projects of the NSFC, one national-level Foreign Experts Project, seven provincial-level research grants, nine teaching-reform projects, and four industry-funded projects.

The Department is committed to training versatile and innovative statisticians equipped with sharp statistical thinking, a broad international outlook, solid methodological foundations, and strong capabilities in practical applications and cross-disciplinary innovation. Over the past five years, nearly 900 professional master’s students in Applied Statistics have been trained on the Zhuhai campus, with more than 610 graduates across four cohorts achieving a 100% employment rate. Current enrollment stands at 150 undergraduates, 284 master's students, and 7 doctoral candidates. For the fall semester of 2025, the Department plans to admit 20 doctoral students, 25 academic master's students, and 304 professional master's students on the Zhuhai campus. Since 2022, undergraduate students have completed 4 national, 5 provincial, and 13 university-level College Student Innovation and Entrepreneurship Training projects, receiving 56 international, 20 national, and 122 provincial awards. Graduate students have secured 20 national and 22 provincial awards under faculty supervision.

北京师范大学文理学院统计系

北京师范大学文理学院为建制性教学科研机构，设立在珠海校区。统计系成立于 2022 年 7 月，是北京师范大学文理学院下设机构。

统计系在珠海校区取得了一系列建设成果，已建成本、硕、博、博士后全层次人才培养体系，获批“统计与数据科学”广东省冲补强重点建设学科（珠海校区首批入选学科之一）和“教育心理与数据科学技术与应用”广东省普通高校重点实验室，并牵头成立粤港澳高校统计与数据科学联盟。当前，统计系面向统计与数据科学国际前沿问题及粤港澳数据科技产业重大战略需求，围绕统计学与计算技术的基础理论、教育大数据的管理与测评方法、复杂结构数据的建模与分析等特色学科方向，全力打造有广泛国际影响力、与北京校区统计学院错位发展、扎根珠海校区的统计与数据科学学科。

统计系现有全职教师 19 人，全部具有博士学位，其中教授 5 人，副教授及副研究员 7 人，讲师 7 人。青年教师全部有海外学习经历，其中 6 人获得海外博士学位。统计系教师在教学和科研上均表现出色，学科带头人朱力行教授是统计学领域首位国家杰出青年科学基金获得者，长江讲座教授，教育部统计学教学指导委员会委员，美国科学促进会、美国统计协会、美国数理统计研究院 Fellow，国际统计学会 Elected Member，并且曾独立获得一些国内，国际的重要的奖项，其中包括国家自然二等奖和德国洪堡研究奖。李勇和张淑梅教授获评北京市高等学校教学名师，获宝钢教育基金优秀教师奖、北京师范大学“四有好老师”金质奖章、北京市教育教学成果二等奖。青年教师谢传龙副教授获评广东省 2024 年度特支计划（科技平台）青年拔尖人才。青年教师不断在 Nature Communications、JASA、Biometrika、IEEE TPAMI、JMLR、JoE 等顶刊发表学术论文，并有青年教师先后获得学校青教赛一等奖、二等奖和彭年杰出青年教师奖。2020 年至今，获批国家自然科学基金重点项目 1 项（珠海校区首个）、面上项目 2 项、青年基金项目 10 项，国家级外专项目 1 项，省部级科研项目 7 项，各级教改项目 9 项，横向项目 4 项。

统计系以培养具有敏锐的统计学思维、开阔的国际视野、扎实的统计学方法论基础、有较强的实践创新应用能力的交叉型拔尖创新统计人才为目标。近五年应用统计专业硕士研究生已在珠海校区培养近 900 人，已毕业四届共 610 余人，毕业生就业率 100%。目前在读本科生 150 人、硕士研究生 284 人、博士研究生 7 人。2025 年秋季学期在珠海校区拟入学博士研究生 20 人、学术型硕士 25 人、专业型硕士 304 人。2022 年至今，指导本科生主持国家级大学生创新训练计划项目 4 项、省级 5 项、校级 13 项，指导本科生获国际级奖励 56 项、国家级奖励 20 项、省级奖励 122 项，指导研究生获得国家级奖励 20 项、省级奖励 22 项。

School of Statistics

Beijing Normal University (BNU) has a distinguished legacy in statistics education, spanning over a century. As early as 1920, the university began offering statistics courses to senior students in the College of Science. Since the founding of the People's Republic of China, BNU's Probability and Mathematical Statistics group—represented by renowned scholars such as Academician Zikun Wang and Professor Shijian Yan—has earned international recognition. The group has cultivated a generation of exceptional educators and leading statisticians who have made significant contributions to the nation.

In 1981, BNU was among the first institutions in China authorized to confer doctoral degrees in Probability Theory and Mathematical Statistics. The discipline was designated a National Key Discipline in 1988 and was recognized as an Innovative Research Group by the National Natural Science Foundation of China in 2002. In 2011, BNU became one of the first universities approved to award doctoral degrees in Statistics as a primary discipline. In 2017, its Statistics program was ranked No. 1 in the ShanghaiRanking's Best Statistics Programs in China.

BNU was also one of the earliest top-tier Chinese universities to establish a dedicated School of Statistics. The School currently offers four undergraduate programs in statistics, leads the development of national first-class undergraduate majors, and has established a national-level virtual teaching and research office. Each year, it enrolls approximately 150 undergraduate and 350 graduate students.

In recent years, the School of Statistics has undertaken 11 major national-level research projects—ranking first nationwide among statistics schools in terms of project volume. Faculty members have published over 40 papers in top-tier journals such as PNAS, Annals of Statistics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society: Series B, and Journal of Econometrics. The School has received four Higher Education Outstanding Scientific Research Output Awards (Science and Technology). In collaboration with the Ministry of Education, the Ministry of Finance, and the National Bureau of Statistics, it has established multiple research platforms, playing a pivotal role in advancing educational evaluation reform and modernizing national statistical systems.

北京师范大学统计学院

北京师范大学统计学教育已有百余年历史，早在 1920 年就面向数理部高年级学生开设了统计学课程。建国以后，以王梓坤院士、严士健先生为代表的概率论与数理统计团队蜚声海内外，为国家培养了一大批优秀统计教师和杰出统计学者。

1981 年，北师大首批获得概率论与数理统计专业博士学位授予权；1988 年，概率论与数理统计被评为国家级重点学科，2002 年，获批国家自然科学基金创新研究群体；2011 年，本学科首批获得统计学一级学科博士学位授予权；2017 年，在软科中国最好统计学科排名中，北师大统计学位列全国第一。

北师大在国内高水平大学中较早成立了统计学院，设有四个统计学类本科专业，承担国家一流本科专业建设任务，建有国家级虚拟教研室，每年招收约 150 名本科生、350 名研究生。

近年来，本学科教师承担了 11 项国家级重大科研项目，数量位居全国统计学科首位。在 PNAS、AoS、Biometrika、JASA、JRSSB、JOE 等顶级期刊上发表论文 40 余篇，获得了四项教育部科学研究优秀成果奖；与教育部、财政部、国家统计局合作建立了多个研究平台，为新时代教育评价改革、国家统计现代化改革做出突出贡献。

Conference Venue

Conference Venue: Beijing Normal University, Zhuhai Campus (BNUZ)

Check-in Location: Lobby of BNUZ International Center

Check-in Time: June 27, from 8:00 AM to 10:00 PM

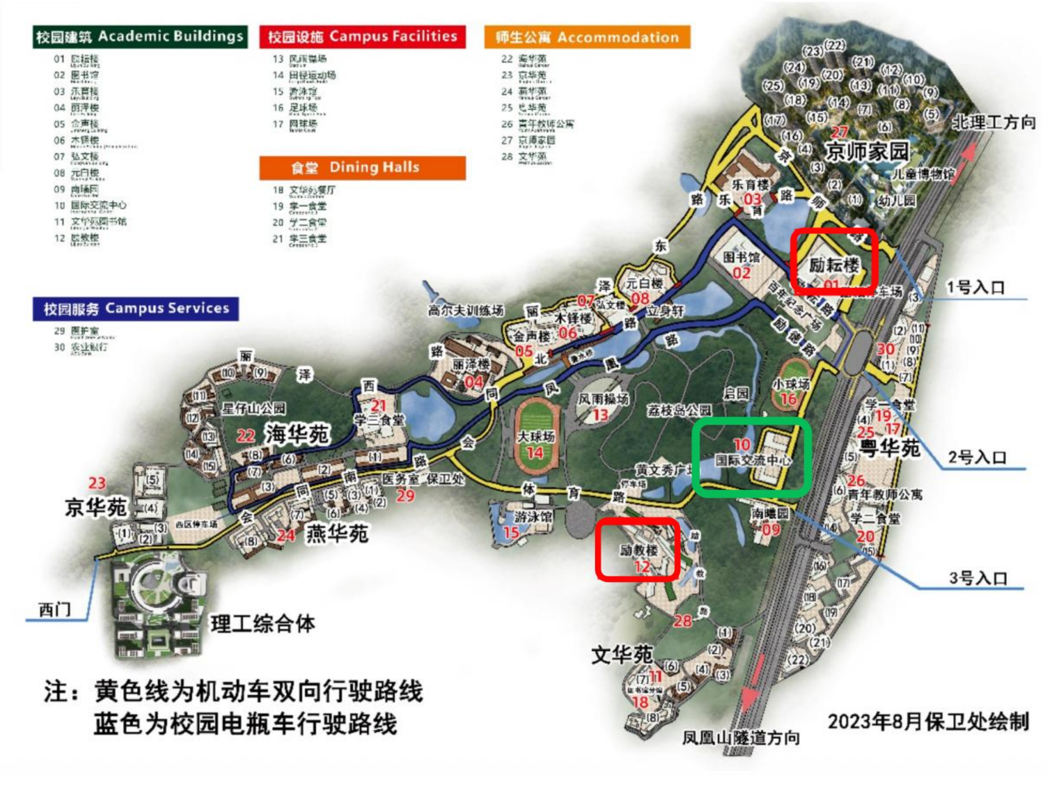
Note: Please proceed to the registration area, which will be arranged alphabetically by Surname, and kindly follow the prompts for on-site registration. Guests who arrive in Zhuhai late on the 27th, please register in the lobby of BNUZ International Center on the morning of the 28th.

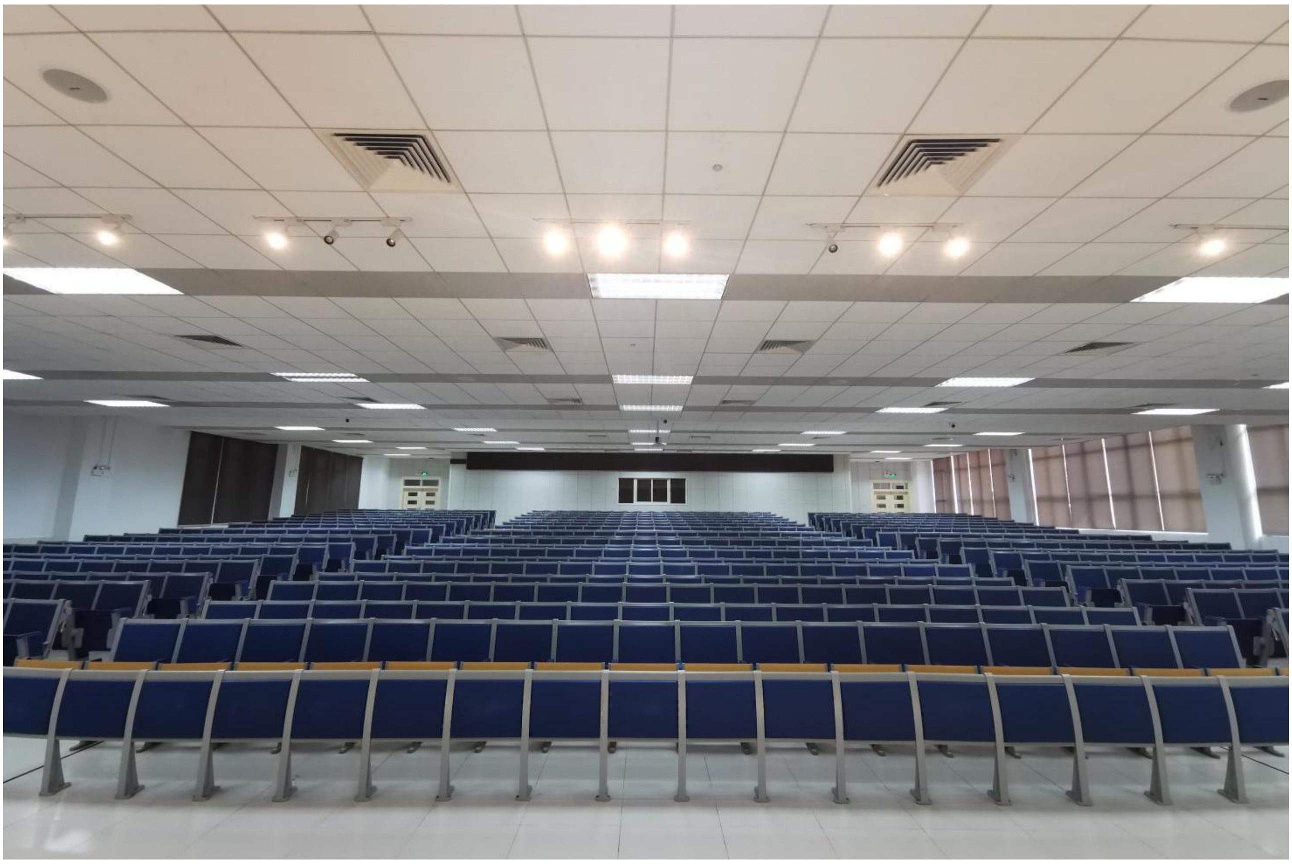
会议地点：北京师范大学珠海校区

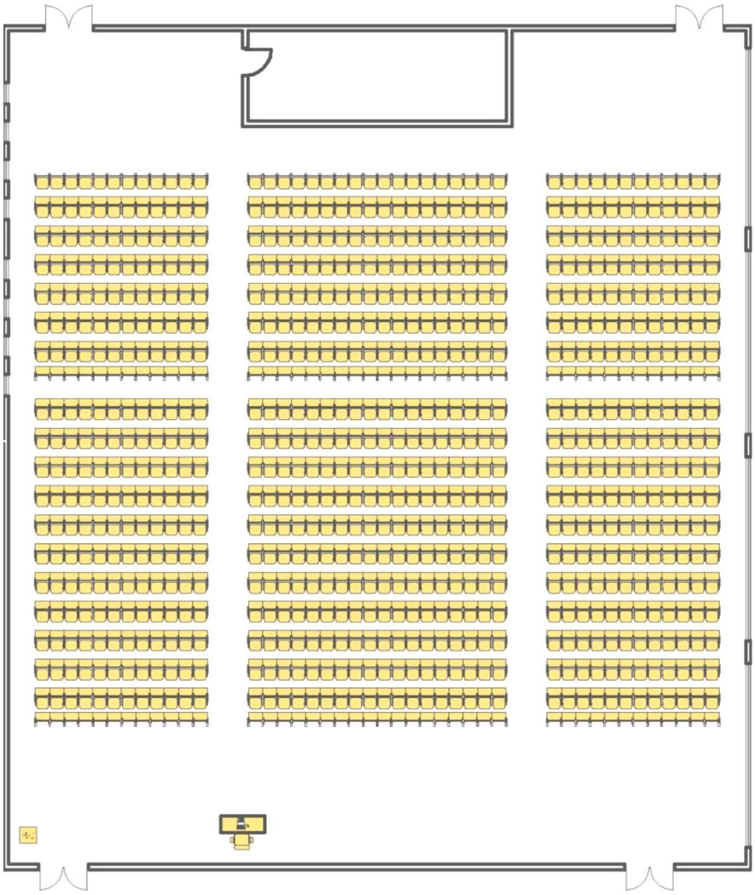
报到地点：北京师范大学珠海校区国际交流中心大堂

报到时间：6 月 27 日 8:00 - 22:00

注：报到现场届时按照姓氏首字母为依据进行分组，请各位嘉宾按照提示进行现场报到；27 日到珠海较晚的嘉宾，请于 28 日上午在国际交流中心大堂报到。

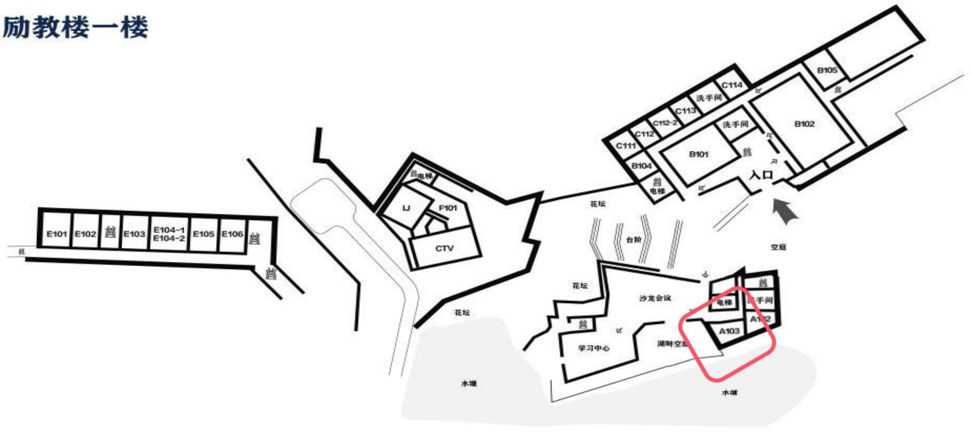




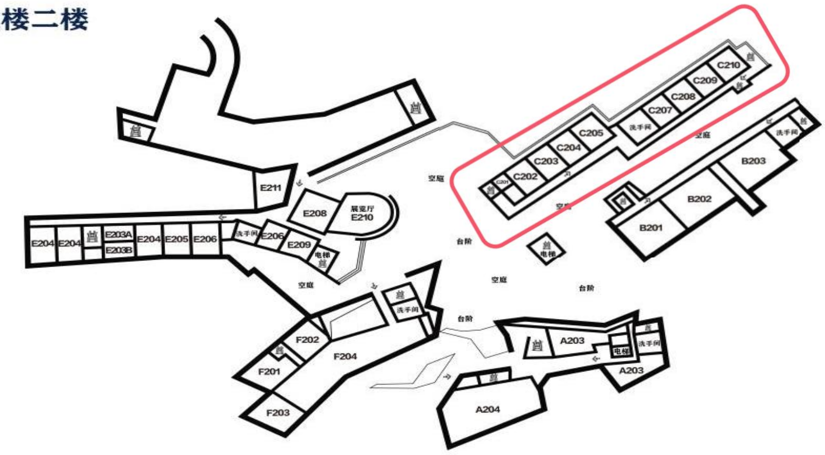


Room A111, Liyun Building

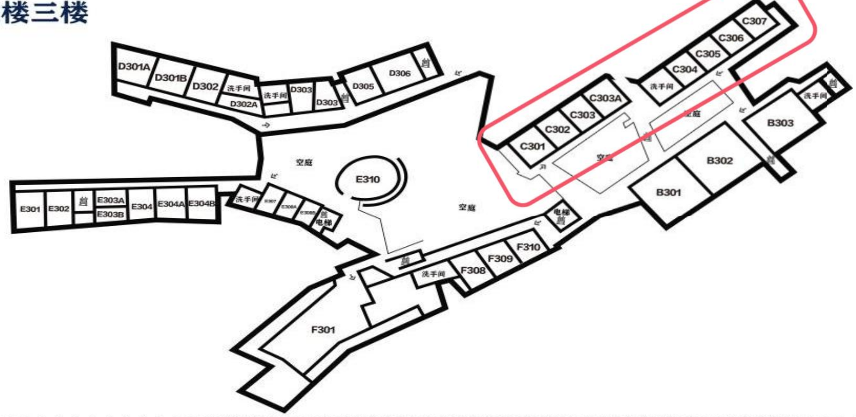
励教楼一楼



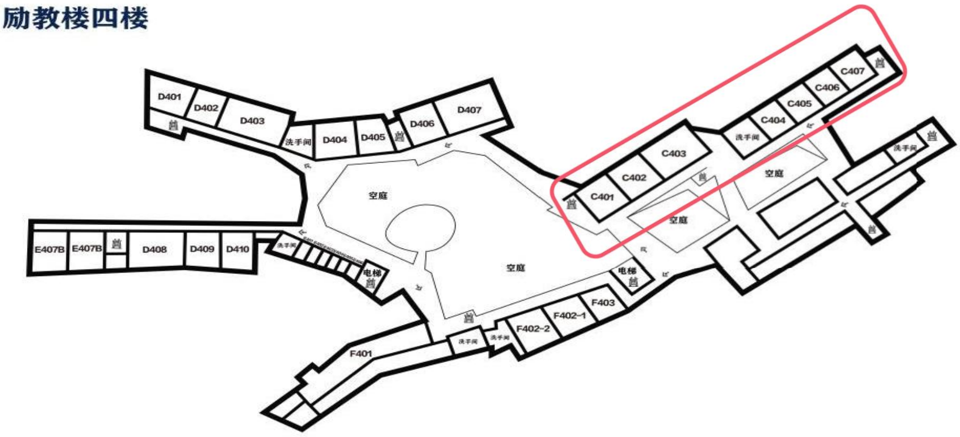
励教楼二楼



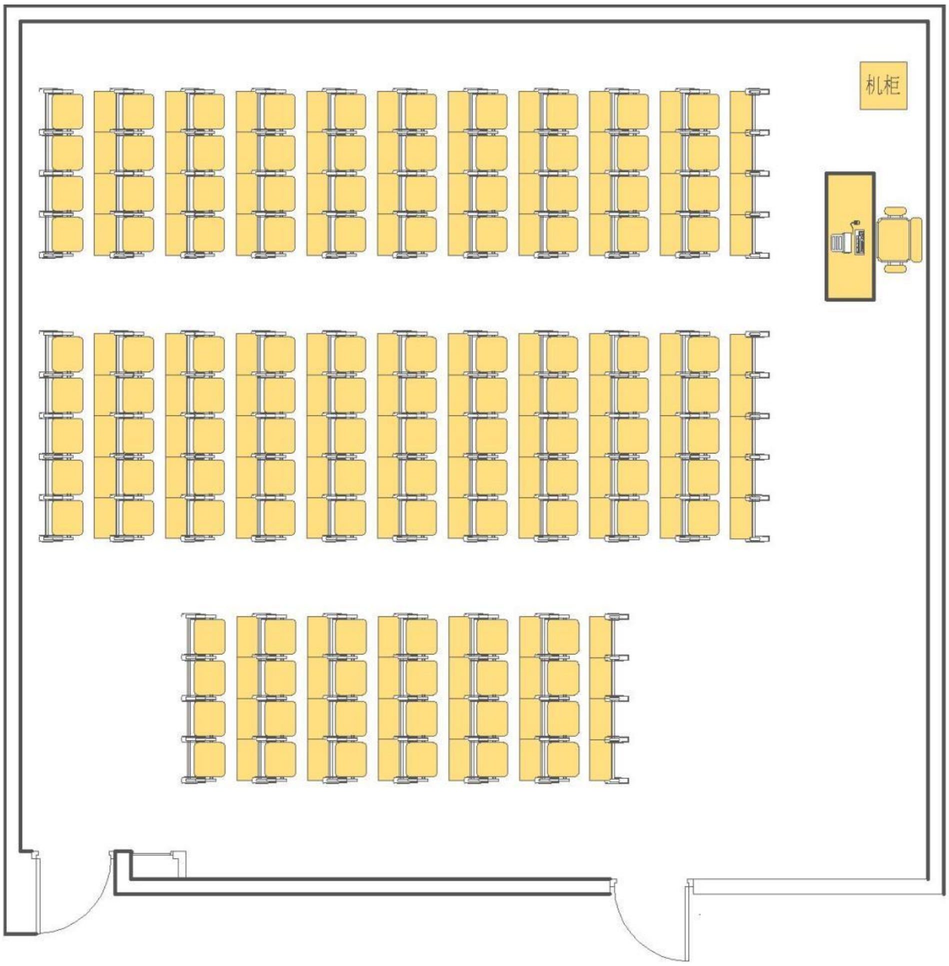
励教楼三楼



励教楼四楼



Lijiao Building



Session room in Lijiao Building

Getting to Beijing Normal University at Zhuhai

1. Via Zhuhai Jinwan Airport

(1) By Zhuhai Airport Express. Purchase tickets via WeChat Official Account "Zhuhai Airport Express" (珠海机场快线). Hotline: 0756-8111333. Options:

a. Business bus: Zhuhai Airport → BNU (Changnanjing Ancient Trail North Bus Stop). Duration: approx. 50 mins | Fare: ¥ 75.

b. Economy Bus: Zhuhai Airport → BNU (Changnanjing Ancient Trail North Bus Stop). Duration: approx. 65 mins | Fare: ¥ 42.

Scan QR code to purchase tickets:

(2) Taxi/Online Ride-hailing. Direct route from Zhuhai Airport to BNU at Zhuhai. Duration: approx. 45 mins | Approx. ¥ 150-200.

2. Via Guangzhou Baiyun Airport

Baiyun Airport Express: Guangzhou Baiyun Airport → Guantang Bus Stop.

Duration: approx.3 hrs | Fare: ¥ 106.

Purchase tickets via WeChat Official Account "Baiyun Airport Express" (白云机场空港快线). Hotline: 020-36063156.

From Guantang Bus Stop to BNU at Zhuhai:

a. Taxi/Online Ride-hailing: Direct to campus. Duration: approx. 10 mins | Approx. ¥ 10-15.

b. Bus: Take B9, B10, 72A, or 70 to Changnanjing Ancient Trail North Bus Stop, then walk approx. 50m to campus. Duration: approx. 26 mins.

Scan QR code to purchase tickets:

3. Via Shenzhen Bao'an Airport

Zhuhai Chimelong Direct Bus: Bao'an Airport → Guantang Bus Stop.

Duration: approx. 60 mins | Fare: ¥ 90.

Purchase tickets via WeChat Official Account "Zhuhai Public Transport & Travel" (珠海公交旅运) → Tap "Custom Routes" → Select "ZhuhaiBao'an Airport".

Hotline: 0756-2116222.

From Guantang Bus Stop to BNU at Zhuhai:

a. Taxi/Online Ride-hailing: Direct to campus. Duration: approx. 10 mins | Approx. ¥ 10-15.

b. Bus: Take B9, B10, 72A, or 70 to Changnanjing Ancient Trail North Bus Stop, then walk approx. 50m to campus. Duration: approx. 26 mins.

Scan QR code to purchase tickets:

4. Via High Speed Rail

(1) Guangzhou South Station: Direct transfer to Zhuhai Tangjiawan Station.
Duration: approx. 60 mins | Fare: ¥ 60.

(2) Zhuhai Mingzhu Station / Zhuhai Station: Direct transfer to Tangjiawan Station.
Duration: 10-20 mins | Fare: ¥ 10.

(3) Tangjiawan Station

From Tangjiawan Station to BNU at Zhuhai:

a. Taxi/Online Ride-hailing: Direct to campus. Duration: approx. 5 mins | Approx. ¥ 5-10.

b. Bus: Take B9, B10, 72A, or 70 to Changnanjing Ancient Trail North Bus Stop, then walk approx. 50m to campus. Duration: approx. 20 mins.

前往北京师范大学珠海校区

1. 经珠海金湾机场

（1）乘坐珠海机场快线，车票可通过扫描“珠海机场快线”购票二维码或线下购买。珠海机场快线咨询电话：0756-8111333，两种车型可供选择：

a. 商务车：珠海机场——北师大（长南迳古道北公交站），车程约50分钟，票价75元；

b. 经济大巴或大型客车：珠海机场——北师大（长南迳古道北公交站），车程约65分钟，票价42元。

以下为“珠海机场快线”购票二维码：

（2）乘坐出租车或网约车由珠海机场直达北京师范大学珠海校区，车程约45分钟，预计150-200元。

2. 经广州白云机场

乘坐白云机场空港快线：广州白云机场——官塘公交站，车程约 3 小时，票价 106 元，车票可通过扫描“白云机场空港快线”购票二维码或线下购买，白云机场空港快线咨询电话：020-36063156。

官塘公交站——北京师范大学珠海校区：

a. 网约车/出租车直达北京师范大学珠海校区，车程约10分钟，预计10-15元；

b. 乘公交车（B9路/B10路/72A路/70路）至长南迳古道北（公交站），再步行约50米至北京师范大学珠海校区，用时约26分钟。

以下为“白云机场空港快线”购票二维码：

2. 经深圳宝安机场

乘坐珠海公交旅运专线：深圳机场汽车客运站——官塘公交站，车程约 60 分钟，票价 90 元。车票可在“珠海公交旅运”公众号购买，关注“珠海公交旅运”，点击左下角“定制线路”，选择“珠海拱北-宝安机场”，随后进入购票页面。珠海公交旅运服务热线：0756-2116222。

官塘公交站——北京师范大学珠海校区：

a. 网约车/出租车直达北京师范大学珠海校区，车程约10分钟，预计10-15元；

b. 乘公交车（B9路/B10路/72A路/70路）至长南迳古道北（公交站），再步行约50米至北京师范大学珠海校区，用时约26分钟。

以下为“珠海公交旅运”二维码：

3. 经高铁

（1）广州南站：同站换乘至珠海唐家湾站，车程约 60 分钟，票价 60 元。

（2）珠海明珠站/珠海站：同站换乘至珠海唐家湾站，车程 10-20 分钟，票价 10 元。

（3）珠海唐家湾站

珠海唐家湾站——北京师范大学珠海校区：

a. 网约车/出租车直达北京师范大学珠海校区，车程约5分钟，预计5-10元；

b. 乘公交车（B9路/B10路/72A路/70路）至长南迳古道北（公交站），再步行约50米至北京师范大学珠海校区，用时约20分钟。

Medical Information

1. Emergency Phone

In case of emergency, please call 120 for medical assistance first. Please give your exact location as clearly as possible (hotel name, conference center name and area, nearby landmarks).

The 24-hour alarm and help number of the Campus Security Office: 0756-3683110, 0756-3621110

Emergency contact of the Local Organization Committee: Zhuo Ren, 0756-3683872, 15245630833.

2. Recommended Hospitals

(1) Medical Office of Beijing Normal University at Zhuhai

Tel: 0756-3621120/3683120.

Address: On Huitong South Road, adjacent to the east side of Building 1, Yanhuayuan.

(2) Zhuhai People's Hospital

Address: No.79 Kangning Road, Xiangzhou District.

Advantages: Large public general Grade A hospital, with complete departments, strong emergency capacity, and foreign medical experience.

(3) The Fifth Affiliated Hospital of Sun Yat-sen University (SASU)

Address: No.52 Meihua East Road, Xiangzhou District.

Advantages: Large public Grade A hospital, relying on Sun Yat-sen University, and with foreign medical experience.

3. Pharmacies

(1) Jialun Guangcai Pharmacy

Address: Haihuayuan bottom commercial, Beijing Normal University at Zhuhai.

(2) Yijia Kang Health Pharmacy

Address: 20 meters west of No.5, Ningtang Village, Jinfeng Road, Xiangzhou District.

It is recommended to carry enough personal prescription drugs for the whole trip, and keep the original packaging and doctor's prescription. Understand China's regulations on imported drugs (especially those containing narcotic and psychotropic ingredients), and avoid carrying prohibited items or excessive amounts.

1.　紧急电话

紧急情况下，请优先拨打 120 寻求医疗急救。尽可能清晰地告知您的具体位置（酒店名称、会议中心名称及区域、附近标志性建筑等）。

学校保卫办 24 小时报警求助电话：0756-3683110、0756-3621110

组委会紧急联系人：任卓（0756-3683872，15245630833）

2.　推荐医院

（1）北京师范大学珠海校区校医务室

电话：0756-3621120/3683120

地址：北京师范大学校内会同南路靠近燕华苑 1 栋东侧。

（2）珠海市人民医院

地址：珠海市香洲区康宁路 79 号。

优势：公立三甲医院，科室齐全，急诊能力强，有涉外医疗经验。

（3）中山大学附属第五医院

地址：珠海市香洲区梅华东路 52 号。

优势：公立三甲医院，依托中山大学，医疗水平高，有涉外医疗经验。

3.　药店

（1）嘉伦光彩大药房

地址：北京师范大学珠海校区海华苑底商。

（2）益嘉康健康药房

地址：珠海市香洲区金凤路宁堂村 5 号正西方向 20 米。

建议携带足够整个行程所需的个人处方药，并保留原包装和医生处方。了解中国对入境药品的规定（尤其含麻醉、精神类成分的药品），避免携带违禁品。

Weather Notice

1. Weather overview

(1) Temperature

The temperature during the day typically ranges from 29°C to 33°C (84°F to 91°F). The nighttime temperature is typically between 25°C and 28°C (77°F and 82°F).

(2) Precipitation

The meeting date is one of the most concentrated periods of rainfall in Zhuhai throughout the year, with almost daily rain possible. The type of rainfall is primarily short, but heavy thunderstorms often occur in the afternoon or evening. Heavy rain or torrential rain may also occur over a more extended period.

(3) Sunshine & Humidity

The sunshine and ultraviolet radiation may be strong, and the humidity is very high.

Suggestion: Keep an eye on the weather forecast to reserve more time for travel to avoid traffic jams on rainy days. Pay attention to heatstroke prevention and cooling. If you have symptoms such as dizziness and nausea, go to a cool place to rest, hydrate, and seek help immediately.

2. Typhoon response

Pay close attention to the official typhoon warning signals (blue, yellow, orange, and red, with red being the highest).

Follow the guidance of the Local Organization Committee and local government.

During the typhoon, stay in a sturdy indoor room away from Windows.

Suggestion: Pay attention to the weather forecast and typhoon warning information released by China Central Meteorological Station, Hong Kong Observatory and weather APPs one week before the meeting and during the meeting.

3. Dressing and Equipment Suggestions

Clothing: Short-sleeved shirts, light and breathable trousers (for mosquito and sun protection), shorts, and skirts. A thin coat or cardigan (to cope with indoor air conditioning that is too cold).

Shoes and socks: Bring several pairs of socks. It is recommended to wear comfortable and breathable walking shoes or waterproof, quick-drying shoes.

Rain gear: Sturdy, folding umbrella or a lightweight raincoat.

Sun protection: High-factor sunscreen, sunglasses, and wide-brimmed hat.

Mosquito repellent: Mosquito repellent liquid or patch (with DEET or Picaridin as the active ingredients, which is more effective).

Others: Portable fan and kettle.

天气须知

1. 天气概况

(1) 温度

日间温度通常在 29°C-33°C (84°F-91°F) 之间，夜间温度通常在 25°C-28°C (77°F-82°F) 之间。

(2) 降水

会议日期处于珠海全年降雨最集中的时段之一，几乎每天都有降雨可能。多为短时强雷阵雨，常在午后或傍晚发生，也可能出现持续时间较长的大雨或暴雨。

(3) 日照与湿度

降雨间隙可能阳光强烈、紫外线辐射强，湿度极高。

建议：随时关注天气预报，出行预留更多时间避开雨天交通拥堵。注意防暑降温，如有头晕、恶心等中暑症状，立即到阴凉处休息、补水、寻求帮助。

2. 台风应对

密切关注官方发布的台风预警信号（蓝色、黄色、橙色、红色，红色为最高）。

遵循会议组织方和当地政府的指引。

避免外出：台风影响期间尽量待在坚固的室内，远离窗户。

建议：会前一周及会议期间，关注中国中央气象台、香港天文台、天气 APP 发布的天气预报和台风预警信息。

3. 着装与装备建议

衣物：短袖衬衫、轻薄透气长裤（防蚊防晒）、短裤、裙子。一件薄外套或开衫（应对室内空调过冷）。

鞋袜：多备几双袜子，建议穿舒适透气的步行鞋或防水、速干鞋。

雨具：推荐携带结实抗风的折叠伞或轻便雨衣。

防晒：高倍数防晒霜、太阳镜、宽檐帽。

防蚊虫：驱蚊液、驱蚊贴，含避蚊胺 DEET 或派卡瑞丁 Picaridin 成分效果较好。

其他：便携小风扇、水壶。

Accommodation

The 2025 ICSA China Conference offers a short-term course scheduled for June 27, with the main conference taking place from June 28 to June 30 at Beijing Normal University Zhuhai Campus. It is recommended that the participants may make room reservation on their own in light of their needs and the organizing committee will recommend the hotels that have signed the preferential price agreements. Nevertheless, there is no guarantee that the agreement price is necessarily the lowest. Please make comparison of the room rates on other platforms and use your discretion. Please refer to the table below for hotel information.

Special reminder: When booking and checking into hotels, please remind the service personnel that we are here to attend the 2025 ICSA China Conference to avoid missing the negotiation price. Please make reservations as early as possible, preferably by phone.

Hotel Information:

	Hotel name	No. of rooms	Contact person	Remark	House price
1	Zhuhai Jinghuayuan Hotel (International Center)	100 rooms	Reservation Phone: 0756-2686999 Email：641160734@qq.com	In the campus	Executive King Room: RMB 428 per room per night; Executive Double Room, Lake View King Room: RMB 388 per room per night
2	Wenhuayuan Apartment	150 rooms	Reservation Phone: 0756-3689666 Email：WHY3689666@qq.com	In the campus	King Room, Standard Double Room: RMB 278 per room per night; Duplex Room: RMB 338 per room per night
3	Tangyi Hotel	150 rooms	Reservation Phone (Yuling Chen):16607561386 Email：1446725630@qq.com	8 kms away from the campus	King Room, Standard Double Room: RMB 448 per room per night
4	Days Inn by Wyndham (Zhuhai Tangjiawan University Town Store)	100 rooms	Reservation Phone: 0756-5652666 or 15767686762 Email：359064226@qq.com	2.4 kms away from the campus	Deluxe King Room, Double Room: RMB 358 per room per night; Scenic View King Room, Double Room: RMB 368 per room per night

住宿

2025 国际泛华统计协会中国会议设有短期课程，短期课程时间为 6 月 27 日，主会议时间为 6 月 28 日- 6 月 30 日，地点为北京师范大学珠海校区。建议参会人员根据需要自行预订酒店，会务组推荐协议酒店，但协议价格并不一定是最低价，请各位参会者对比其他平台的价格后作出最有利于自己的选择。酒店信息请参考下表。

特别提醒：参会人员预订和入住酒店时，请提示服务人员是来参加 2025 国际泛华统计协会中国会议的，以免错失协议价。请与会人员尽早预定，最好通过电话预定。

酒店信息：

	酒店名称	房间数量	联系人	备注	房 价
1	国交中心—珠海京华苑大酒店	100	订房电话：0756-2686999 邮箱：641160734@qq.com	校内	行政大床房 428 元/间/天; 行政双人房、湖影大床房 388 元/间/天
2	文华苑公寓	150	订房电话：0756-3689666 邮箱: WHY3689666@qq.com	校内	大床房、标准双人房 278 元/间/天; 复式房间 338 元/间/天
3	唐邑酒店	150	订房电话：16607561386 邮箱： 1446725630@qq.com	距学校 8 公里	大床房、双人房 448 元/间/天
4	戴斯精选温德姆酒店	100	订房电话：0756-5652666, 15767686762 邮箱：359064226@qq.com	距学校 2.4 公里	雅致大床房、双人房 358 元/间/天; 景观大床房、双人房 368 元/间/天

Banquet

In order to facilitate informal discussions among participants, complimentary meals and tea breaks will be provided.

Dining location & Times

Venue: International Center, 1F Western Restaurant & 2F Chinese Restaurant

Date	Meal	Time	Location
June 28	Lunch	12:00-14:00	International Center
	Dinner	17:30-19:30	International Center
June 29	Lunch	11:30-13:30	International Center
	Banquet	18:30-21:00	International Center
June 30	Lunch	11:30-13:30	International Center

Conference Banquet

The conference banquet will be held on June 29 (Sunday) night, Western Restaurant on the 1st Floor and Chinese Restaurant on the 2nd Floor of the International Center.

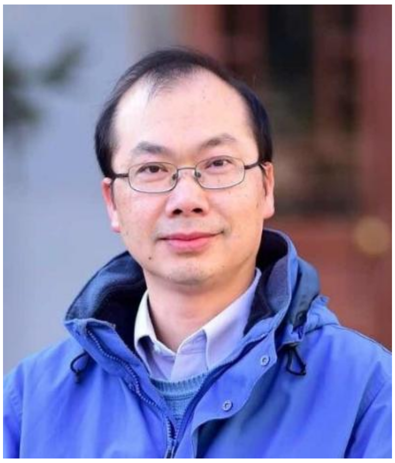
The banquet will start at 18:30.

Banquet speaker: *Dr. Tian Zheng*

The tea breaks will be held in the lobbies of the conference buildings.

Program Overview

Day 1 (June 28, 2025)			
	In person front desk open from 7:30am-7:30pm		Host
9:30AM-10:00AM	Room A111 Liyun Building	Welcome and Opening Ceremony	Lixing Zhu
10:00AM-11:00AM		Keynote Lecture 1—Runze Li	Yuanjia Wang
11:00AM-11:20AM	Coffee Break		
11:20AM-12:20PM	Room A111 Liyun Building	Keynote Lecture 2—Wenguang Sun	Yuanjia Wang
12:20AM-2:00PM	International Center	Lunch	
2:00PM-3:40PM	Lijiao Building	Parallel Invited Sessions	Session Chairs
3:40PM-4:00PM	Coffee Break		
4:00PM-5:40PM	Lijiao Building	Parallel Invited Sessions	Session Chairs
5:30PM-7:30PM	International Center	Dinner	
Day 2 (June 29, 2025)			
	In person front desk open from 7:30am-7:30pm		Host
8:30AM-10:10AM	Lijiao Building	Parallel Invited Sessions	Session Chairs
10:10AM-10:30AM	Coffee Break		
10:30AM-12:10PM	Lijiao Building	Parallel Invited Sessions	Session Chairs
12:10AM-1:30PM	International Center	Lunch	
2:00PM-3:40PM	Lijiao Building	Parallel Invited and Contributed Sessions	Session Chairs
3:40PM-4:00PM	Coffee Break		
4:00PM-5:40PM	Lijiao Building	Parallel Invited and Contributed Sessions	Session Chairs
6:30PM-9:00PM	International Center	Conference Banquet—Tian Zheng	Hongyu Zhao
Day 3 (June 30, 2025)			
	In person front desk open from 7:30am-2:00pm		Host
8:30AM-10:10AM	Lijiao Building	Parallel Invited and Contributed Sessions	Session Chairs
10:10AM-10:30AM	Coffee Break		
10:30AM-12:10PM	Lijiao Building	Parallel Invited and Contributed Sessions	Session Chairs
12:10AM-1:30PM	International Center	Lunch	

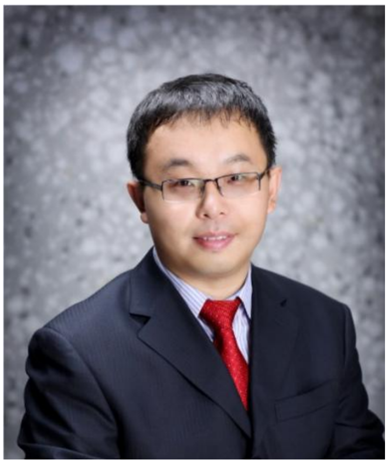


Runze Li, Eberly Family Chair Professor in Statistics at Pennsylvania State University. He served as Co-Editor of Annals of Statistics from 2013 to 2015. Runze Li is Fellow of IMS, ASA and AAAS. His recent honors and awards also include the Distinguished Achievement Award of International Chinese Statistical Association, 2017, Faculty Research Recognition Awards for Outstanding Collaborative Research. College of Medicine, Penn State University in 2018 and Distinguished Mentoring Award, Eberly College of Science, Penn State University in 2023. His research interests include theory and methodology in variable selection, feature screening, robust statistics, nonparametric and semiparameteric regression. His interdisciplinary research aims to promote the better use of statistics in social behavioral research, neural science research and climate studies.

Time: 10:00 AM to 11:00 AM, June 28, 2025

Statistical Inference in High-Dimensional Linear and Generalized Linear Models

Abstract: This talk will begin with an overview of recent development of statistical inference in high-dimensional linear and generalized linear models, and then focus on a double power-enhanced testing procedure for inference on high-dimensional linear hypotheses in high-dimensional regression models. Through a projection approach that aims to separate useful inferential information from the nuisance one, the newly proposed test accurately accounts for the impact of high-dimensional nuisance parameters. With a carefully-designed projection matrix, the projection procedure enables us to transform the problem of interest into a test on moment conditions, from which we construct a U-statistic-based test that is applicable in simultaneous inference on a diverging number of linear hypotheses. We prove that under regularity conditions, the plug-in test statistic converges to its oracle counterpart, acting as well as if the nuisance parameters were known in advance. Asymptotic null normality is established to provide convenient tools for statistical inference, accompanied by rigorous power analysis. To further strengthen the testing power, we develop two power enhancement techniques to boost the power from two distinct aspects respectively, and integrate them into one powerful testing procedure to achieve double power enhancement. The finite-sample performance is demonstrated using simulation studies, and an empirical analysis of a real data example.



Wenguang Sun, Professor and Doctoral Supervisor at Zhejiang University. The current director of the Data Science. Research Center at Zhejiang University. Bachelor's degree from Peking University in 2003. In 2008, he obtained his PhD from the University of Pennsylvania under the guidance of renowned statistician and COPSS award winner Professor Cai Tianwen. Wenguang Sun's main research interests include large-scale statistical inference, integrated analysis and transfer learning, conformal prediction and inference, empirical Bayesian methods, and statistical decision theory. Before returning to China, he served as a tenured full professor at the Marshall School of Business at the University of Southern California (USC) in the United States. He has been selected for the National High level Talent Program, won the Frontiers of Science Award at the International Basic Science Conference, the CAREER Award from the American Science Foundation, and has been awarded the USC Business School Outstanding Research Award and the Golden Apple Best Teaching Award multiple times. Received funding from the National Science Foundation of the United States as the principal investigator (PI) four times from 2010 to 2022. He once served as the editorial board member of authoritative academic journals such as Scientia Sinica Mathematica, Journal of the Royal Statistical Society–Series B, and Journal of Multivariate Analysis.

Time: 11:20 AM to 12:20 PM, June 28, 2025

Title: Conformal Meets Empirical Bayes: A New Perspective on Distribution-Free Multiple Testing

Abstract: This talk presents two novel approaches that integrate conformal inference with empirical Bayes techniques for distribution-free multiple testing. The first approach, Conformalized Locally Adaptive Weighting (CLAW), addresses the multiple testing problem in the presence of side information. By constructing pairwise exchangeable scores and leveraging a mirror process, CLAW achieves finite-sample control of the false discovery rate (FDR) while exploiting structural information in auxiliary covariates to enhance power. The second approach tackles the common challenge of unknown null distributions in large-scale testing. We introduce a conformalized multiple testing method that utilizes self-calibrated empirical null samples (SENS), thereby avoiding reliance on estimated null distributions. Rather than depending on stringent assumptions or asymptotic approximations, SENS requires only a mild symmetry condition on the error distribution, delivering finite-sample FDR control while achieving asymptotic optimality under weak conditions. Together, these methods demonstrate a powerful synergy between conformal inference and empirical Bayes principles. By unifying mirror processes, pairwise exchangeability, and adaptive score construction, CLAW and SENS pursue both optimality and robustness to enable efficient decision-making across a range of scenarios. This research is joint with my PhD students, Zinan Zhao and Yang Tian, at Zhejiang University.

Banquet Talk

Tian Zheng, Professor and Department Chair of Statistics at Columbia University. She obtained her Ph.D. from Columbia in 2002. In her research, she develops novel methods for exploring and understanding patterns in complex data from different application domains such as biology, psychology, climatology, and etc. Her current projects are in the fields of statistical machine learning, spatiotemporal modeling, and social network analysis, collaborating with ecologists and earth scientists. Professor Zheng's research has been recognized by the 2008 Outstanding Statistical Application Award from the American Statistical Association (ASA), the Mitchell Prize from ISBA, and a Google research award. She became a Fellow of the American Statistical Association in 2014. Professor Zheng is passionate about education and mentoring. From 2015-2016, she was one of the series creators for Columbia's edX Massive Online Open Course (MOOC) series on data science. From 2017-2020, she was associate director for education of Columbia Data Science Institute. She led a number of education programs, including the MS in Data Science program at Columbia, data science capstone projects with data ethics components, DSI Scholars program that connects students with academic research projects in data science, the Collaboratory program for interdisciplinary data science curriculum development, a number of popular Data Science boot camps. She created DSI's working group on Data Science Education and has been coordinating data science education efforts across Columbia. Professor Zheng is the receipt of the 2017 Columbia's Presidential Award for Outstanding Teaching. In 2021, she was recognized by a Lenfest Distinguished Columbia Faculty Award that recognizes the excellence of faculty as teachers and mentors of both undergraduate and graduate students.

Time: 6:30 PM to 8:30 PM, June 29, 2025

Title: Statistics is What Statisticians Do

Abstract: Statistics gains significance through the impacts it creates in collaborative applications. In this talk, I will highlight a few examples from my own professional journey to illustrate how statisticians enhance scientific rigor and foster data-driven innovations. Through case studies across multiple disciplines, I will demonstrate that the essence of statistics lies not merely in techniques, but fundamentally in how statisticians engage with complex, real-world problems to make meaningful contributions.

题目：统计之道：统计学家之所为

摘要：统计学的价值往往体现在统计学家通过合作应用而实现的实际影响力上。在这个演讲里，我会分享一些经历与合作案例，展示统计学家如何推动科学严谨性以及基于数据的创新。透过多个学科的实例，我将阐述统计学的本质不仅仅在于方法和理论本身，而在于统计学家如何深入实际问题，与其他领域密切协作，从而创造真正有意义的成果。

Junior Researcher Award Winners

Jiaqi Huang, Beijing Normal University

Title: Dimension-reduction and detection for structurally changed tensor sequence

Time: June 29th AM, 10:30-10:55

Session 25CHI115: Junior Researcher Award Winners Session

Room: C406

Cheng Yu, Tsinghua University

Title: Two-way Matrix Autoregressive Model with Thresholds

Time: June 29th AM, 10:55-11:20

Session 25CHI115: Junior Researcher Award Winners Session

Room: C406

Yuqian Zhang, Renmin University of China

Title: Dynamic treatment effects: high-dimensional doubly robust inference under model misspecification

Time: June 29th AM, 11:20-11:45

Session 25CHI115: Junior Researcher Award Winners Session

Room: C406

Dingke Tang, Washington University in St. Louis

Title: The synthetic instrument: From sparse association to sparse causation

Time: June 29th AM, 11:45-12:10

Session 25CHI115: Junior Researcher Award Winners Session

Room: C406

Haobo Zhang, Tsinghua University

Title: Sup-norm convergence rate and uniform inference for kernel ridge regression

Yuexin Chen, Renmin University of China

Title: Projective Kernel Two-Sample Test in High Dimension

Lingxuan Shao, Fudan University

Title: Ordinary Differential Equation Models for a Collection of Discretized Functions

a. June 28th PM (14:00-15:40)

Number	Title	Organizer	Chair	Room
25CHI004	Advanced Statistical Methods and Applications in Medical and Image Data Analysis	Jinhan Xie	Jinhan Xie	C203
25CHI005	Advanced Statistical Methods for analyzing health-related data.	Liping Zhu	Jin Liu	C204
25CHI009	Advances in Complex Time Series and Spatial Modelling and Learning	Zudi Lu	Zudi Lu	C205
25CHI012	Advances in Statistical Learning and Algorithm	Lu Lin	Lu Lin	C207
25CHI016	Advances in Statistical Methods for Complex Data Integration and Causal Inference	Chenglong Ye	Xiaofang Shi	C208
25CHI022	Complex Structured Data Analysis	Ling Zhou	Ling Zhou	A103
25CHI023	Contemporary survival analysis and new applications	Yichuan Zhao	Yichuan Zhao	C209
25CHI026	Efficient data collection and computing techniques in data-rich era	Mingyao Ai	Jun Yu	C210
25CHI033	Innovations in Network Analysis	Chenlei Leng	Wenlin Dai	C301
25CHI036	Innovative Approaches in Statistical Modeling and Analysis	Jinfeng Xu	Jinfeng Xu	C302
25CHI038	Innovative methodology and strategy in statistical analysis	Joyce Wang	Joyce Wang	C303
25CHI040	Innovative Statistical Learning Methods and Applications	Xinyuan Song	Xiangnan Feng	C304
25CHI042	Kernel methods in machine learning	Qian Lin	Qian Lin	C305
25CHI047	Modeling average and related topics	Hua Liang	Hua Liang	C306
25CHI050	Modern Statistical Inference for Complex Data	Xianyang	Guanxun Li	C404
25CHI052	Modern statistical methods in biostatistics	Yanyuan Ma	Wenbin Lu	C307
25CHI064	Recent Advance in High-dimensional Modelling	Yingxing Li	Yingxing Li	C405
25CHI091	Semi-parametric and nonparametric methods for complex data analysis	Peijun Sang	Peijun Sang	C406

b. June 28th PM (16:00-17:40)

Number	Title	Organizer	Chair	Room
25CHI070	Recent advances in network modeling	Junhui Wang	Junhui Wang	C203
25CHI080	Recent developments in analyzing complex data	Chenlei Leng	Guodong Li	C204
25CHI087	Recent Statistical Advances in Complex Genetic and Genomic Data Analysis	Yuehua Cui	Yuehua Cui	C205
25CHI089	Sample Size, Power, and Likelihood	Penny Peng	Gengsheng Qin	C207
25CHI092	Some important topics in pharmaceutical statistics	Yixin Fang	Baoying Yang	A103

25CHI094	Specific Statistical considerations in clinical trial design	Ning Li	Ning Li	C208
25CHI097	Statistical analyses of several types of complex data	Fei Chen	Fei Chen	C404
25CHI099	Statistical Inference on high-dimensional covariance matrix	Shurong Zheng	Shurong Zheng	C209
25CHI101	Statistical Learning and Medical Diagnostics	Ngai Hang	Jinfeng Xu	C210
25CHI105	Statistical methods elevated by modern computation and massive data	Yanyuan Ma	Jiwei Zhao	C301
25CHI106	Statistical Methods for Survival Data with Complex Censoring and Missing or Mismeasured Covariates	Yanqing Sun	Yanqing Sun	C302
25CHI108	Statistical Methods in Medical Applications	Jialiang Li	Jialiang Li	C303
25CHI112	Structured machine learning	Junhui Wang	Ben Dai	C304
25CHI114	Transforming clinical trials with causal inference thinking and methodology	Zhiwei Zhang	Min Zhang	C305
25CHI001	Advanced Experimental Design and Subsampling Approaches for Complex Data Analysis	Mingyao Ai	Yaping Wang	C405
25CHI002	Advanced Learning Methods for Complex Medical Data	Hua Liang	Hua Liang	C306
25CHI006	Advanced Statistical Methods for Spatial Transcriptomics	Liping Zhu	Qing Cheng	C406
25CHI007	Advancements in High-Dimensional Statistical Methods and Applications	Jinfeng Xu	Zhenggang Wang	C307

c. June 29th AM (8:30-10:10)

Number	Title	Organizer	Chair	Room
25CHI008	Advances in Causal Discovery for Omics Data	Shanghong Xie	Haoran Xue	C203
25CHI011	Advances in Spatial Statistics with Random Field Modeling: Methods, Metrics, and Applications	Juan Du	Juan Du	C204
25CHI014	Advances in Statistical Methods and Applications	Lizhe Sun	Lizhe Sun	C205
25CHI017	Advances in Statistical Modeling: Variable Selection, Dependence, and Nonparametric Methods	Hongmei Jiang	Hongmei Jiang	C207
25CHI021	Causal inference and decision-making	Yifan Cui	Yuanshan Gao	C208
25CHI027	Frontier Statistical Methods for Single-cell RNA Sequencing Data	Xiaodan Fan	Xiaodan Fan	C209
25CHI030	Innovations and Partnerships in Data-Rich Environments: Emerging Advances in Matrix and Tensor Modeling	Jiangyan Wang	Jiangyan Wang	C404
25CHI041	Integrate Statistics into Deep Learning for Digital Image Processing and Analysis	Weihong Guo	Junying Meng	C210
25CHI044	Machie learning for data assimilation	Yuling Jiao	Shuyi Zhang	C301
25CHI046	Modeling and inference for distributions and high dimensional data	Ming-Yen Cheng	Jialiang Li	C302
25CHI059	Novel Machine Learning Methods for Disease Progression and Precision Medicine	Shanghong Xie	Huichen Zhu	C304
25CHI061	Observational data analysis with complex study designs	Andy Ni	Yuzi Zhang	C305

25CHI063	Random matrices and high-dimensional statistics	Jeff Yao	Jeff Yao	C306
25CHI069	Recent Advances in Microbiome Data Analysis	Tao Wang	Tiantian Liu	C405
25CHI073	Recent Advances in Single-cell Data Analysis	Tao Wang	Tao Wang	C307

d. June 29th AM (10:30-12:10)

Number	Title	Organizer	Chair	Room
25CHI075	Recent Advances in Statistical Machine Learning: Theory and Algorithms	Yunwen Lei	Yunwen Lei	C203
25CHI076	Recent Advances on the Analysis of Failure Time Data	Jianguo Sun	Jianguo Sun	C204
25CHI078	Recent development in statistical methods for related regression models and applications	Zhiqiang Cao	Jie He	C205
25CHI079	Recent developments about high-dimensional inference	Xu Guo	Long Feng	C207
25CHI081	Recent developments in causal inference and survival analysis	Yifan Cui	Yuanshan Gao	C208
25CHI082	Recent Developments in Covariate Adjustment for Randomized Clinical Trials	Xin Zhang	Xin Zhang	C209
25CHI083	Recent Developments in High-dimensional Data Analysis	Xingqiu Zhao	Xingqiu Zhao	C210
25CHI084	Recent developments in reinforcement learning and mobile health	Yifan Cui	Tao Shen	C301
25CHI088	Robust Analysis for Treatment Decision and Risk Prediction under Complex Data Settings	Donglin Zeng	Yu Gu	C302
25CHI093	Some recent developments about big data analysis	Xu Guo	Lei Wang	C304
25CHI096	Statistical Advances in Large Language Models and Network Analysis	Will Wei Sun	Will Wei Sun	C305
25CHI098	Statistical analysis of complex survival data	Chunjie Wang	Shuying Wang	C306
25CHI111	Statistics for Emerging Trends in Machine Learning	Yafei Wang	Yafei Wang	C307
25CHI024	Design and analysis of clinical studies	Zhezhen Jin	Zexi Cai	C404
25CHI065	Recent Advances in Complex Data	Yang Zhou	Yang Zhou	C405
25CHI115	Junior Researcher Award Winners Session	---	Xingche Guo	C406

e. June 29th PM (14:00-15:40)

Number	Title	Organizer	Chair	Room
25CHI003	Advanced Statistical and Computational Methods for Microbiome and Metagenomics Data Analysis	Yanan Zhao	Jiyuan Hu	C203
25CHI010	Advances in Modern Statistical Methodologies: Robust Estimation, High-Dimensional Inference, and Innovative Biomedical Applications	Xiaoya Xu	Xiaoya Xu	C204
25CHI018	Advancing Multi-platform and Multi-modal Omics Harmonization	Qian Li	Wei Liu	C404

25CHI019	Advancing Multi-Regional Clinical Trials: Methodology and Application Considerations for Ensuring Global Representation, Regulatory Harmonization, and Ethical Integrity	Menggang Yu	Menggang Yu	C405
25CHI025	Dimension Reduction Methods	Xin (Henry) Zhang	Ning Wang	C205
25CHI028	High dimensional statistics inference	Guangming Pan	Guangming Pan	C207
25CHI029	High dimensional statistics inference (II)	Guangming Pan	Bo Zhang	C208
25CHI031	Innovations in Causal Inference and Statistical Methods for Complex Data Structures	Yidong Zhou	Doudou Zhou	C209
25CHI034	Innovations in Nonparametric and Functional Data	Xingche Guo	Xingche Guo	C210
25CHI043	Lifetime Data Analysis	Mei-Ling Ting Lee	Mei-Ling Ting Lee	C301
25CHI045	Model fairness and challenges and development of statistical models	Zhezhen Jin	Yingwei Paul Peng	C302
25CHI051	Modern Statistical Learning	Lexin Li	Yin Xia	C304
25CHI056	New advances in statistical theory, method and application	Le Zhou	Le Zhou	C307
25CHI057	New frontiers in large-scale data analysis with applications to heterogeneous data	Yichuan Zhao	Yichuan Zhao	C305
25CHI062	On Statistical Stability and Ensemble Learning	Wei Zhong	Wei Zhong	C306

f. June 29th PM (16:00-17:40)

Number	Title	Organizer	Chair	Room
25CHI066	Recent Advances in Correcting Measurement Error in Epidemiological Research	Xin Zhou	Xin Zhou	C404
25CHI067	Recent advances in high dimensional data and machine	Xiaochao Xia	Xiaochao Xia	C203
25CHI068	Recent Advances in Machine Learning Techniques for Point Process Models	Biao Cai	Shizhe Chen	C204
25CHI074	Recent Advances in Statistical and Machine Learning	Chengchun Shi	Jin Zhu	C205
25CHI085	Recent developments of high dimensional model checking	Falong Tan	Falong Tan	C405
25CHI086	Recent Progresses in Nonparametric and Semiparametric Statistics	Ying Yan	Ying Yan	C207
25CHI090	Scalable Learning and Knowledge Transfer for Complex Biomedical Data	Xinping Cui	Gang Li	C208
25CHI095	Statistical Advances for Integrative Multi-Omics Data Analysis	Jiebiao Wang	Jiebiao Wang	C209
25CHI102	Statistical learning based on high dimensional and complex data	Ming-Yen Cheng	Jin-Ting Zhang	C210
25CHI109	Statistical Network Analysis and Applications	Binyan Jiang	Binyan Jiang	C301
25CHI110	Statistical theory of neural networks	Jun Fan	Jun Fan	C302

25CHI113	Theoretical Advances in Machine Learning, Dimension Reduction, and Functional Data Analysis	Dongming Huang	Dongming Huang	C304
25CHI037	Innovative Inference Methods for Complex Data: Bridging Theory and Practice	Xingche Guo	Xingche Guo	C305
25CHI048	Modern Machine Learning: Tackling Real-World Data Challenges	Xinping Cui	Xiaoqian Liu	C306
25CHI049	Modern Multivariate Analysis for Tensor and Multiview Data	Xin Zhang	Jing Zeng	C307

g. June 30th AM (8:30-10:10)

Number	Title	Organizer	Chair	Room
25CHI013	Advances in Statistical Learning and Network Analysis for Complex Data	Peng Liu	Xiaofei Zhang	C203
25CHI020	Advancing Risk Management with Statistical Learning	Fan Yang	Fan Yang	C204
25CHI032	Innovations in High Dimensional Complex Data Analysis: From Functional Data Analysis to Measurement Error Modeling	Juan Xiong	Juan Xiong	C207
25CHI035	Innovative Approaches in Electronic Health Record (EHR) Data Analysis	Molei Liu	Molei Liu	C208
25CHI055	New advances in design and analysis of longitudinal studies	Zhigang Li	Zhigang Li	C209
25CHI071	Recent Advances in Nonparametric Estimation and Inference	Qing Wang	Qing Wang	C210
25CHI072	Recent advances in privacy-protected data collection and analysis	Samuel Wu	Zhigang Li	C302
25CHI103	Statistical Learning for Complex Data Structures	Ruiyang Wu	Yuchen Zhou	C304
25CHI104	Statistical learning with complex data	Peter Song	Xinyuan Song	C305

h. June 30th AM (10:30-12:10)

Number	Title	Organizer	Chair	Room
25CHI039	Innovative Statistical and Machine Learning Methods for Complex Health Data	Jingjing Zou	Todd Ogden	C203
25CHI053	Modern Statistical Modeling in Medical Research with Real World Data	Zheyu Wang	Jing Huang	C204
25CHI054	New Advancements in Statistical Learning	Zheyu Wang	Yaofang Hu	C207
25CHI058	New statistical methods in nonlinear regression analyses	Peter Song	Xuerong Chen	C208
25CHI060	Novel statistical methods for complex data analysis	Yichuan Zhao	Yichuan Zhao	C209
25CHI077	Recent Advences in Statistical Learning for Biological and Biomedical Data	Ruiyang Wu	Ruiyang Wu	C210
25CHI107	Statistical methods for the analysis of complex data	Zhezhen Jin	Antai Wang	C302

Name	Affiliation	Title
Group 1: Bayesian Methods and Applications, June 29th PM, 14:00-15:40, C406 **Chair:** Zhe Fei, University of California, Riverside		
Hengtao Zhang	Guangdong Ocean University	Bayesian analysis of Cox-type regression model with partly linear covariate effects via reversible jump Markov chain Monte Carlo
Mingan Yang	University of New Mexico	Bayesian crossover trial with binary data and extension to Latin-square design
Nanwei Wang	University of New Brunswick	Birth-Death MCMC for Bayesian Variable Selection in Polygenic Risk Score Models
Zhihao Wu	The Chinese University of Hong Kong	Tree-Based Bayesian Methods for Analyzing Partially Ordered Latent Statuses of Parkinson's Disease
Group 2: High-Dimensional Inference, June 29th PM, 16:00-17:40, C406 **Chair:** Yu Guo, The Chinese University of Hong Kong		
Bin Liu	Fudan University	A General U-Statistic Framework for High-Dimensional Multiple Change-Point Analysis
Lingfeng Lyu	University of science and technology of China	Quadratic form estimation for moderate-dimensional logistic regression models
Yuexin Chen	Renmin University of China	Randomized empirical likelihood test for ultra-high dimensional means under general covariances
Yu Guo	The Chinese University of Hong Kong	Debiased Inference for High-Dimensional Censored Quantile Regression via L1 Penalization
Group 3: Statistical Tests & Diagnostics, June 30th AM, 8:30-10:10, C306 **Chair:** Chuhan Wang, Beijing Normal University		
Haiming Lin	Zunyi Normal University; Guangdong University of Finance & Economics	Significance test and application of principal components
Shaoyun Zhang	Shanghai University of International Business and Economics	Influence Diagnostics for Generalized CP Tensor Regression Models
Peiwen Jia	Peking University	Smooth Tests for Normality in ANOVA
Yuxin Tao	Southern University of Science and Technology	Homogeneity pursuit in ranking inference based on pairwise comparison
Group 4: Causal Inference & Treatment Effects, June 30th AM, 8:30-10:10, C307 **Chair:** Qiang Zhao, Northeast Normal University		
Xinyi Xu	Ohio State University	Double-Score Gaussian Process Model for Robust Causal Inference in Observational Studies
Qixian Zhong	Xiamen University	Deep Orthogonal Learner for Conditional Quantile Treatment Effect Estimation
Xintong Li	East China Normal University	Efficient Semi-supervised Estimation of Optimal Individualized Treatment Regime with Survival Outcome

Qiang Zhao	Northeast Normal University	Optimal Designs for Order-of-Addition Two-Level Factorial Experiments
Group 5: Network & Graph Analysis, June 30th AM, 10:30-12:10, C304 **Chair:** Yanni Zhang, Beijing Normal University		
Qian Hui	Fudan University	Systemic Risk Management via Maximum Independent Set in Extremal Dependence Networks
Xiyue Zhu	University of Science and Technology of China	Testing Global Community Structure in Multi-Layer Networks: A Leave-One-Out Polynomial Statistic
Yinqiao Yan	Beijing University of Technology	Spatially aware adjusted Rand index for evaluating spatial transcriptomics clustering
Yunhe Pan	The University of New South Wales	Rate Guarantees for recovery of latent space distances
Group 6: Time Series & Functional Data, June 30th AM, 10:30-12:10, C305 **Chair:** Sicheng Fong, The Chinese University of Hong Kong		
Sicheng Fong	The Chinese University of Hong Kong	Optimal Subsampling and EM Algorithms for Non-Markovian Semiparametric Regression with Interval-Censored Multi-State Data
Zerui Guo	Sun Yat-sen University	A Unified Principal Component Analysis for Stationary Functional Time Series
Qirui Hu	Tsinghua University	From sparse to dense functional time series: phase transitions of detecting structural breaks and beyond
Cheng Yu	Tsinghua University	Two-way Matrix Autoregressive Model with Thresholds
Group 7: Biostatistics & Medical Applications, June 30th AM, 10:30-12:10, C306 **Chair:** Hua He, Tulane University		
Yuyang He	The Chinese University of Hong Kong	Mixed membership latent variable model with unknown factors, factor loadings and number of extreme profiles
Hua He	Tulane University	Joint Modeling Approach for censored predictors in generalized linear model due to detection limit with applications to metabolites data
Mengyu Li	Renmin University of China	Double Optimal Transport for Differential Gene Regulatory Network Inference with Unpaired Samples
Fangyi Wei	The University of Hong Kong	Nested Deep Learning Model Towards a Foundation Model for Brain Signal Data
Group 8: Dimension Reduction & Variable Selection, June 30th AM, 10:30-12:10, C307 **Chair:** Pengfei Wang, Nanyang Technological University		
Zhengtian Zhu	Chinese Academy of Sciences	False Discovery Control for High-Dimensional Linear Models with Model-X Knockoff and p-values
Meggie Wen	Missouri University of Science and Technology	Multi-Population Sufficient Dimension Reduction
Xin Zhou	University of Science and Technology of China	SOFARI-R: High-Dimensional Manifold-Based Inference for Latent Responses
Pengfei Wang	Nanyang Technological University	Overview of normal-reference tests for high-dimensional means with implementation in the R package ‘HDNRA’

Time: 4:00 PM to 6:00 PM, June 28, 2025

Venue: The first-floor hall of Area B, Lijiao Building

No.	Name	Affiliation	Title
1	Chaodong Chen	Harbin Institute of Technology, Shenzhen	Anomaly Detection in Multivariate Time Series Based on Residual Distribution Using a Hybrid Model Combining Bayesian LSTM and Graph Model
2	Chao Deng	Shanghai Jiao Tong University	Bridging unpaired single-cell multimodal data for integrative analyses with SuperMap
3	Mingxuan Ge	University of Michigan	Revisiting matching prior to diff-in-diff in the presence of non-parallel trends
4	Zijian Huang	Harbin Institute of Technology, Shenzhen	Stochastic Approximation MM Algorithms for Multiple Responses Mixed-effects Model and its Application
5	Daoyuan Lai	The University of Hong Kong	Bayesian Transfer Learning for Enhanced Estimation and Inference
6	Xiang Li	The University of Hong Kong	A Bayesian fine-mapping model using a continuous global-local shrinkage prior with applications in prostate cancer analysis
7	Zhifei Li	Beijing Normal University	SELF-Tree: An Interpretable Model for Multivariate Causal Direction Heterogeneity Analysis
8	Ziyu Liu	Weill Cornell Medicine	Dirichlet process mixture model for Optimizing Grouping to Enhance power in clinical trials
9	Jun Liu	Georgia Southern University	Detecting Olympic Tourism Legacy using Structural Change Tests
10	Ruoxi Lyu	The University of Hong Kong	Evaluating time-varying and heterogeneous efficacy of COVID-19 antiviral drugs
11	Haochen Rao	Beijing Normal University	Estimation and Inference for Density-convoluted Support Vector Machine with Streaming Data
12	Kang Shuai	Peking University	Identification and estimation of causal peer effects using instrumental variables
13	Peng Wang	Jilin Univeristy	DisPRP: A Novel Criterion for Assessing Replicability through Distinguishability
14	Moshu Xu	Tsinghua University	Simultaneous inference for eigensystems of functional time series with application
15	Jiajing Xue	Xiamen University	Ordinal Sparse Neural Network in Interaction Analysis
16	Jie Zhang	Indiana University School of Medicine	Identify the link between genetic factor and normal tissue morphological heterogeneity through association study
17	Houlin Zhou	Anhui University	Change-point detection in stochastic differential equations

a. June 28th PM (14:00-15:40)

25CHI004: Advanced Statistical Methods and Applications in Medical and Image Data Analysis
Room: C203
Organizer: Jinhan Xie
Chair: Jinhan Xie

14:00 High dimensional proteomics data added value in heart failure patient phenomapping

♦Yinggan Zheng, Cindy Westerhout, Paul Armstrong

University of Alberta, University of Alberta, University of Alberta

14:25 Variational Bayesian Logistic Tensor Regression with Application to Image Recognition

Yunzhi Jin, ♦Yanqing Zhang, Niansheng Tang

Yunnan University, Yunnan University, Yunnan University

14:50 Conditional inference for ultrahigh-dimensional additive hazards model

♦Meiling Hao

University of International Business and Economics

15:15 Tensor-Based Individualized Treatment Rules for Neuroimaging Applications

♦Yang Sui, Yuanying Chen, Ting Li, Yang Bai, Hongtu Zhu

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, The University of North Carolina at Chapel Hill

25CHI005: Advanced Statistical Methods for analyzing health-related data.
Room: C204
Organizer: Liping Zhu
Chair: Jin Liu

14:00 High-dimensional covariate-augmented overdispersed poisson factor model

Wei Liu, ♦Qingzhi Zhong

Sichuan University, Jinan University

14:25 Heterogeneous change-point Effects in Longitudinal Data: An Application to Age-Related Cognitive Decline

Xiaoke Li, Boxian Wei, ♦Guangyu Yang, Min Zhang

Vanke School of Public Health, Tsinghua University, Vanke School of Public Health, Tsinghua University, Institute of Statistics and Big Data, Renmin University of China, Vanke School of Public Health, Tsinghua University

14:50 Multivariable Mendelian Randomization Method accounting for complex correlated and uncorrelated pleiotropy

♦Qing Cheng, Li Cao

Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics

Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics

15:15 Statistical Inference with Mixed-Effect Model for Covariate-Adaptive Randomized Experiments

♦Yang Liu, Lucy Xia, Feifang Hu

Renmin University of China, Hong Kong University of Science and Technology, The George Washington University

25CHI009: Advances in Complex Time Series and Spatial Modelling and Learning
Room: C205
Organizer: Zudi Lu
Chair: Zudi Lu

14:00 Coefficient Shape Transfer Learning for Functional Linear Regression

♦Shuhao Jiao, Ian Mckeague, Ngai-Hang Chan

City University of Hong Kong, Columbia University, City University of Hong Kong

14:25 Nonparametric Estimation of Weakly Dependent Time Series via Neural Networks

Zudi Lu, Shubin Wu, ♦Gan Yuan, Chao Zheng

City University of Hong Kong, University of Southampton, City University of Hong Kong, University of Southampton

14:50 A Root-n-Consistent Semiparametric Superquantile Autoregression for Dynamic Time Series with a Possibly Incorrect Model Specification

Jiangtao Wang, ♦Zudi Lu, Xiyu Zhou, Wu Jin

School of Economics and Business Administration, Central China Normal University, China, Department of Biostatistics, City University of Hong Kong, Hong Kong SAR, China, School of Economics and Business Administration, Central China Normal University, China, School of Economics and Business Administration, Central China Normal University, China

15:15 Covariance parameter estimation for spatial models

♦Saifei Sun

City University of Hong Kong

25CHI012: Advances in Statistical Learning and Algorithm
Room: C207

Organizer: Lu Lin
Chair: Lu Lin

14:00 Optimal Model Averaging for Imbalanced Classification

♦*Ze Chen, Jun Liao, Wangli Xu, Yuhong Yang*

Shandong University, Renmin University of China, Renmin University of China, Tsinghua University

14:25 A paradox in Metropolis-Hastings practice

♦*Jiandong Shi, Sheng Lian, Xiaodan Fan*

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

14:50 The Binary and Ternary Quantization Can Improve Feature Discrimination

♦*Weiyu Li, Weizhi Lu, Mingrui Chen*

Shandong University, Shandong University, Shandong University

15:15 Quantile-Matched DC in Massive Data Regression

Yan Chen, ♦*Lu Lin*

Shandong University, Shandong University

25CHI016: Advances in Statistical Methods for Complex Data Integration and Causal Inference
Room: C208
Organizer: Chenglong Ye
Chair: Xiaofang Shi

14:00 Deep Clustering Evaluation: How to Validate Internal Clustering Validation Measures

Zeya Wang, ♦*Chenglong Ye*

University of Kentucky, University of Kentucky

14:25 Robust High-dimensional Inference for Causal Effects Under Unmeasured Confounding and Invalid Instruments with an Application to Multivariable Mendelian Randomization Analysis

♦*Yunan Wu, Lan Wang, Baolin Wu, Yixuan Ye, Hongyu Zhao*

Tsinghua University, University of Miami, UC Irvine, Yale University, Yale University

14:50 High-Resolution Feature Identification in High-Dimensional Clustering

♦*Lyuou Zhang*

Shanghai University of Finance and Economics

15:15 Combining variable screening methods for model averaging in High-Dimensional Data Analysis

♦*Zhihao Zhao, Yuhong Yang, Li Wen*

Capital University of Economics and Business, Tsinghua University, Renmin University of China

25CHI022: Complex Structured Data Analysis
Room: A103
Organizer: Ling Zhou
Chair: Ling Zhou

14:00 High-dimensional large-scale mixed-type data imputation under missing at random

♦*Wei Liu, Guizhen Li, Ling Zhou, Lan Luo*

School of Mathematics, Sichuan University, School of Economics and Finance, Guizhou University of Commerce, Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Department of Biostatistics and Epidemiology, Rutgers University

14:25 Identification of Latent Subgroups for Time-varying Panel Data Models

♦*Ye He, Qing Luo, Liu Liu, Shengzhi Mao, Ling Zhou*

Sichuan Normal University, Sichuan Normal University, Chengdu University of Technology, Southwestern University of Finance and Economics, Southwestern University of Finance and Economics

14:50 A Functional Semiparametric Mixed Effects State Space Model with Prior Information for County Level Spatiotemporal Data

♦*Mengying You, Wensheng Guo*

Shanghai University of International Business and Economics, University of Pennsylvania

25CHI023: Contemporary survival analysis and new applications
Room: C209
Organizer: Yichuan Zhao
Chair: Yichuan Zhao

14:00 SurGAN: A Generative Adversarial Network Model for Tabular Survival Data

♦*Hong Wang*

Central South University

14:25 A model-free correlation coefficient for censored data

♦*Linlin Dai, Tengfei Li, Kani Chen*

Southwestern University of Finance and Economics, University of North Carolina at Chapel Hill, Hong Kong University of Science and Technology

14:50 Fiducial inference in survival analysis

♦*Yifan Cui*

Zhejiang University

15:15 Nonparametric estimation of a state entry time distribution conditional on a (past) state occupation using current status data.

Samuel Anyaso-Samuel, ♦*Somnath Datta*

NIH, U of Florida

25CHI026: Efficient data collection and computing techniques in data-rich era
Room: C210
Organizer: Mingyao Ai
Chair: Jun Yu

14:00 Maximum projection Latin hypercube designs using number theoretic methods

Yuxing Ye, Ru Yuan, ♦Yaping Wang

East China Normal University, Zhongnan University of Economics and Law, East China Normal University

14:25 Data-driven Sampling Based Stochastic Gradient Descent Method

Yanjing Feng, Shiqi Zhou, ♦Yongdao Zhou

Nankai University, Nankai University, Nankai University

14:50 A Wasserstein distance-based spectral clustering method for transaction data analysis

♦Yingqiu Zhu, Danyang Huang, Bo Zhang

University of International Business and Economics, Renmin University of China, Renmin University of China

15:15 BanditSIS: efficient algorithm for large-sample feature screening via multi-armed bandits

♦Cheng Meng

Renmin University of China

25CHI033: Innovations in Network Analysis
Room: C301
Organizer: Chenlei Leng
Chair: Wenlin Dai

14:00 Temporal network analysis via a degree-corrected Cox model

Yuguo Chen, Lianqiang Qu, Jinfeng Xu, ♦Ting Yan, Yunpeng Zhou

University of Illinois at Urbana-Champaign, Central China Normal University, City University of Hong Kong, Central China Normal University, The University of Hong Kong

14:25 A community Hawkes model for continuous-time networks with interaction heterogeneity

Haosheng Shi, ♦Wenlin Dai

Renmin University of China, Renmin University of China

14:50 Modeling reciprocity in directed networks

♦Rui Feng, Chenlei Leng

University of Warwick, University of Warwick

15:15 A Network Propagation Model for Graph-linked Data

♦Yingying Ma, Chenlei Leng

Beihang University, School of Economics and Management, University of Warwick

25CHI036: Innovative Approaches in Statistical Modeling and Analysis
Room: C302
Organizer: Jinfeng Xu
Chair: Jinfeng Xu

14:00 Ratio-controlled screening for structural break predictive regressions

♦Rongmao Zhang, Zhenjie Qin, Yang Zu

Zhejiang Gongshang University, Zhejiang University, University of Macau

14:25 Modeling paired binary data by a new bivariate Bernoulli model with flexible beta kernel correlation

Xunjian Li, Shuang Li, Guo-Liang Tian, ♦Jianhua Shi

Department of Statistics and Data Science, Southern University of Science and Technology, Department of Mathematics, Dongguan University of Technology, Department of Statistics and Data Science, Southern University of Science and Technology, School of Mathematics and Statistics, Minnan Normal University

14:50 Flexible DNA Methylation Analysis scBS-Seq Data]{scFMA: A Flexible Random Effects Model for DNA Methylation Analysis with scBS-Seq Data

♦Yanting Wu, Xifen Huang, Yao Lu, Jinfeng Xu, Hengjian Cui

Yunnan Normal University

15:15 Model free feature screening for large scale and ultrahigh dimensional survival data

Yingli Pan, Haoyu Wang, ♦Zhan Liu

Hubei University, Hubei University, Hubei University

25CHI038: Innovative methodology and strategy in statistical analysis
Room: C303
Organizer: Joyce Wang
Chair: Joyce Wang

14:00 Application of Bayesian hierarchical model for subgroup analysis in vaccine efficacy study

♦Joyce Wang

Sanofi

14:25 Missing data matter: an empirical evaluation of the impacts of missing EHR data in comparative effectiveness research

♦Yizhao Zhou, Jiasheng Shi, Ronen Stein, Xiaokang Liu, Robert Baldassano, Christopher Forrest, Yong Chen, Jing Huang

Department of Biometrics, China, Astrazeneca, Inc.

14:50 Integration of central statistical surveillance into central statistical surveillance (for binary endpoint)

♦Xiaojia Zhang

Sanofi

25CHI040: Innovative Statistical Learning Methods and Applications
Room: C304
Organizer: Xinyuan Song
Chair: Xiangnan Feng

14:00 Transformed dynamic quantile regression for case-cohort studies

♦*Haijin He*

Shenzhen University

14:25 Kernel Density Balancing with Application in Hi-C data

♦*Ning Hao*

The University of Arizona

14:50 Bayesian Inference of Phenotypic Plasticity of Cancer Cells Based on Dynamic Model for Temporal Cell Proportion Data

Shuli Chen, Yuman Wang, Da Zhou, ♦*Jie Hu*

Xiamen University

15:15 A new non-parametric resampling method based on representative points

♦*Sirao Wang, Yinan Li, Kai-Tai Fang , Huajun Ye*

Hong Kong Baptist University, Hong Kong Baptist University, Beijing Normal-Hong Kong Baptist University (BNBU), Beijing Normal-Hong Kong Baptist University (BNBU)

25CHI042: Kernel methods in machine learning
Room: C305
Organizer: Qian Lin
Chair: Qian Lin

14:00 On the Pinsker bound of inner product kernel regression in large dimensions

♦*Weihao Lu, Jialin Ding, Haobo Zhang, Qian Lin*

National University of Singapore, Tsinghua University, Tsinghua University, Tsinghua University

14:25 Diffusion Actor-Critic: Formulating Constrained Policy Iteration as Diffusion Noise Regression for Offline Reinforcement Learning

♦*Wenjia Wang*

The Hong Kong University of Science and Technology (Guangzhou)

14:50 Nonparametric Estimation of Mixed MNLs by Kernel Machine

♦*Liang Ding*

Fudan University

15:15 On non-redundant and linear operator-based nonlinear dimension reduction

Zhoufu Ye, ♦*Wei Luo*

Zhejiang University, Zhejiang University

25CHI047: Modeling average and related topics
Room: C306
Organizer: Hua Liang
Chair: Hua Liang

14:00 A Subsampling Strategy for AIC-based Model Averaging with Generalized Linear Models

Jun Yu, ♦*HaiYing Wang, Mingyao Ai*

Beijing Institute of Technology, University of Connecticut, Peking University

14:25 PEARL: Performance-enhanced Aggregated Representation Learning

♦*Wenhui Li, Shijing Gong, Xinyu Zhang*

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, School of Management, University of Science and Technology of China, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

14:50 Functional sufficient dimension reduction with multivariate responses: A projection averaging method and beyond

♦*Wenchao Xu*

Shanghai University of International Business and Economics

15:15 Quantile Regression Model Averaging for Distributed Data

♦*Haili Zhang*

Shenzhen Polytechnic University

25CHI050: Modern Statistical Inference for Complex Data
Room: C404
Organizer: Xianyang Zhang
Chair: Guanxun Li

14:00 Transfer Learning for Survival Data Using Pseudo Observations

♦*Hanxuan Ye*

University of Pennsylvania

14:25 Adaptive Independence Test via the Generalized HSIC

♦*Yaowu Zhang*

Shanghai University of Finance and Economics

14:50 Statistical Inference for Differentially Private Stochastic Gradient Descent

♦*Zhanrui Cai, Xintao Xia, Linjun Zhang*

The University of Hong Kong, Iowa State University, Rutgers University

15:15 Fast Association Recovery in High Dimensions by Parallel Learning

Ruipeng Dong, ♦*Canhong Wen*

University of Science and Technology of China, University of Science and Technology of China

25CHI052: Modern statistical methods in biostatistics
Room: C307
Organizer: Yanyuan Ma
Chair: Wenbin Lu

14:00 Self-Consistent Equation-guided Neural Networks for Censored Time-to-Event Data

Sehwan Kim, Rui Wang, ♦*Wenbin Lu*

Ewha Womans University, Department of Population

Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Department of Statistics, North Carolina State University

14:25 A Multimodal Functional Deep Learning Approach for Multi-omics Data

Yuan Zhou, ♦Pei Geng, Shan Zhang, Feifei Xiao, Guoshuai Cai, Li Chen, Qing Lu

University of Florida, University of New Hampshire, Michigan State University, University of Florida, University of Florida, University of Florida, University of Florida

14:50 A new time-varying coefficients regression model for predicting COVID-19 deaths

♦Juxin Liu, Brandon Bellows, Joan Hu, Jianhong Wu, Zhou Zhou, Chris Soteros, Lin Wang

University of Saskatchewan, University of Saskatchewan, Simon Fraser University, York University, University of Toronto, University of Saskatchewan, University of New Brunswick

15:15 Assessing Algorithm Fairness Requires Adjustment for Risk Distribution Differences Across Population Subgroups: A Unified Framework for Fairness Evaluation

♦Xiaoyi Zheng, Hong Zhang, Sarah Hegarty, Jinbo Chen

Department of Statistics and Finance, University of Science and Technology of China, Anhui, China, Department of Statistics and Finance, University of Science and Technology of China, Anhui, China, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

25CHI064: Recent Advance in High-dimensional Modelling
Room: C405
Organizer: Yingxing Li
Chair: Yingxing Li

14:00 Testing High-Dimensional Effects in Quantile Regression with High-Dimensional Confounding: A Decorrelated Smoothing Approach

♦Peirong Xu

Shanghai Jiao Tong University

14:25 The Inferential Theory of Random Projection Methods

♦Hongjun Li

Tsinghua University

14:50 Low-rank and sparse network regression

Aureo de Paula, ♦Yingxing Li, Weining Wan

University College London, Xiamen University, University of Bristol

15:15 Estimation of Large Dynamic Precision Matrices with a Latent Semiparametric Structure

♦Jia Chen, Yuning Li, Oliver Linton

University of Macau, University of York, University of Cambridge

25CHI091: Semi-parametric and nonparametric methods for complex data analysis
Room: C406
Organizer: Peijun Sang
Chair: Peijun Sang

14:00 Statistical methods for transfer learning in survival analysis

♦Yu Gu, Donglin Zeng, Danyu Lin

University of Hong Kong, University of Michigan, University of North Carolina at Chapel Hill

14:25 Spatial deconvolution and cell type-specific spatially variable gene detection in spatial transcriptomics

♦Yuehua Cui

Michigan State University

14:50 Checking the Cox Proportional Hazards Model with Interval-Censored Data

♦Yangjianchen Xu, Donglin Zeng, Danyu Lin

University of Waterloo, University of Michigan, University of North Carolina at Chapel Hill

15:15 Distributional Off-Policy Evaluation with Deep Quantile Process Regression

♦Fan Zhou

Shanghai University of Finance and Economics

b. June 28th PM (16:00-17:40)

25CHI070: Recent advances in network modeling
Room: C203
Organizer: Junhui Wang
Chair: Junhui Wang

16:00 Moment-integrated Bias-adjusted Spectral Method for Community Detection in Multi-layer Networks

♦Xuefei Wang, Junhui Wang, Gaorong Li

School of Statistics, Beijing Normal University, Department of Statistics, The Chinese University of Hong Kong, School of Statistics, Beijing Normal University

16:25 Data Integration: Network-Guided Covariate Selection in High-Dimensional Data

♦Wanjie Wang, Tao Shen

National University of Singapore, National University of Singapore

16:50 A dynamic network autoregressive model for time-varying network-link data

♦Jingnan Zhang, Bo Zhang, Yu Chen

University of Science and Technology of China, University of Science and Technology of China, University of Science and Technology of China

17:15 False Discovery Rate Control Using Bi-Gaussian Mirrors

♦Binyan Jiang

The Hong Kong Polytechnic University

25CHI080: Recent developments in analyzing complex data
Room: C204
Organizer: Chenlei Leng
Chair: Guodong Li

16:00 A Unified Analysis of Likelihood-based Estimators in the Plackett-Luce Model

♦Ruijian Han, Yiming Xu

The Hong Kong Polytechnic University, University of Kentucky

16:25 Approximation Error from Discretizations and Its Applications

♦Junlong Zhao, Xiumin Liu, Bin Du, Yufeng Liu

Beijing Normal University, Beijing Normal University, Beijing Normal University, University of North Carolina at Chapel Hill

16:50 Large-Scale Curve Time Series with Common Stochastic Trends

♦Degui Li, Yuning Li, Peter C.B. Phillips

University of Macau, University of York, Yale University

17:15 Supervised Factor Modeling for High-Dimensional Linear Time Series

♦Guodong Li

University of Hong Kong

25CHI087: Recent Statistical Advances in Complex Genetic and Genomic Data Analysis
Room: C205
Organizer: Yuehua Cui
Chair: Yuehua Cui

16:00 Differential Inference for Single-cell RNA-Sequencing Data

Fangda Song, Kevin Yip, ♦Yingying Wei

The Chinese University of Hong Kong, Shenzhen, Sanford Burnham Prebys Medical Discovery Institute, The Chinese University of Hong Kong

16:25 Genetic association testing with multivariate survival phenotypes under interval censoring

Juhee Lee, ♦Chenxi Li, Gongjun Xu, Qing Lu

Michigan State University, Michigan State University, University fo Michigan, University of Florida

16:50 A unified framework for identification of cell-type-specific spatially variable genes in spatial transcriptomic studies

Zhiwei Wang, Yeqin Zeng, Ziyue Tan, Yuheng Chen, Xinrui Huang, Hongyu Zhao, Zhixiang Lin, ♦Can Yang

HKUST, HKUST, HKUST, HKUST, HKUST, Yale, CUHK, HKUST

17:15 Hypothesis testing in high-dimensional censored-transformation models

Xiao Zhang, Xiangyong Tan, Runze Li, ♦Xu Liu

The Chinese University of Hong Kong, Shenzhen, Jiangxi University of Finance and Economics, Pennsylvania State University, Shanghai University of Finance and Economics

25CHI089: Sample Size, Power, and Likelihood
Room: C207
Organizer: Penny Peng
Chair: Gensheng Qin

16:00 Influence Function-based Empirical Likelihood for AUC in Presence of Covariates

Baoying Yang, Xinjie Hu, ♦Gengsheng Qin

Southwest Jiaotong University, CHINA, Georgia State University, USA, Georgia State University, USA

16:25 Optimizing Sample Size in vaccine efficacy trial: integrating the timing expectation of interim analysis and the seasonallity prediction of diseases

♦Penny Peng

Department of Biostatistics and Programming, China, Sanofi, Inc.

16:50 The Identifiability of Copula Models for Dependent Competing Risks Data With Exponentially Distributed Margins

♦Antai Wang

New Jersey Institute of Technology

17:15 Considering the correlation between serotypes for the sample size estimation in vaccine clinical trials

♦Jieqi Jin, Fabrice Bailleux, Jian Ding, Ian Deng

Department of Biostatistics and Programming, China, Sanofi, Inc.

25CHI092: Some important topics in pharmaceutical statistics
Room: A103
Organizer: Yixin Fang
Chair: Baoying Yang

16:00 Multiple Comparisons Procedures for Analyses of Joint Primary Endpoints and Secondary Endpoints

Xiaolong Luo, Lerong Li, Oleksandr Savenkov, Weijian Liu, Xiao Ni, Weihua Tang, ♦Wenge Guo

Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, New Jersey Institute of Technology

16:25 Matching-Assisted Power Prior for Incorporating Real-World Data in Clinical Trials

Ruoyuan Qian, Biqing Yang, Xinyi Xu, ♦Bo Lu

The Ohio State University, The Ohio State University, The Ohio State University, The Ohio State University

16:50 Consistency consideration for a region in a multiple regional clinical trial

♦En-Tzu (Angela) Tang

Abbvie Inc.

17:15 Sequential Monitoring of Covariate Adaptive Randomized Clinical Trials with Nonparametric Approaches

Xiaotian Chen, Jun Yu, ♦Hongjian Zhu, Li Wang

AbbVie Inc., AbbVie Inc., Systlmmune Inc., AbbVie Inc.

25CHI094: Specific Statistical considerations in clinical trial design
Room: C208
Organizer: Ning Li
Chair: Ning Li

16:00 Statistical consideration of estimands in vaccine clinical trials

♦Yufan Deng

Sanofi China

16:25 A Bayesian phase I/II platform design with survival efficacy endpoint for dose optimization

Xian Shi, Jin Xu, ♦Rongji Mu

East China Normal University, East China Normal University, Shanghai Jiao Tong University

16:50 An overview of regional treatment effect evaluation via information borrowing in MRCTs

♦Yanghui Liu

Sanofi

17:15 Timeline prediction in event driven clincial trials

♦Zhini Wang

Sanofi

25CHI097: Statistical analyses of several types of complex data
Room: C404
Organizer: Fei Chen
Chair: Fei Chen

16:00 Copula-based models in compositional data analysis

Caikun Chen, Yu Fei, ♦Pengyi Liu, Zhuo Chen

Department of Statistics, Yunnan University of Finance and Economics

16:25 Growth curves mixture model for longitudinal data based on mean–covariance modeling

♦Yating Pan, Fangfang Pan, Jianxin Pan

Yunnan University of Finance and Economics, Yunnan University of Finance and Economics, Beijing Normal University, BNU-HKBU United International College

16:50 Unified optimal model averaging with a general loss function based on cross-validation

♦Dalei Yu, Xinyu Zhang, Hua Liang

Xi'an Jiaotong University, University of Science and Technology of China and Academy of Mathematics and Systems Science, Chinese Academy of Sciences, George Washington University

25CHI099: Statistical Inference on high-dimensional covariance matrix
Room: C209
Organizer: Shurong Zheng
Chair: Shurong Zheng

16:00 High-dimensional scale invariant discriminant analysis

♦Ming Li, Cheng Wang, Yanqing Yin, Shurong Zheng

Shandong Technology and Business University

16:25 Sparse estimation of high-dimensional cross-covariance matrices and its applications

♦Kazuyoshi Yata, Tetsuya Umino, Makoto Aoshima

University of Tsukuba, University of Tsukuba, University of Tsukuba

16:50 Testing for large-dimensional covariance matrix under differential privacy

Shiwei Sang, ♦Yicheng Zeng, Shurong Zheng, Xuehu Zhu

Xi'an Jiaotong University, Sun Yat-sen University, Northeast Normal University, Xi'an Jiaotong University

17:15 Approximate Normality in testing hierarchical covariance structures belonging to a quadratic subspace

Daniel Klein, ♦Yuli Liang, Mateusz John

P. J. Safarik University in Kosice, Slovakia, Guangxi Normal University, China, Institute of Mathematics, Poznan University of Technology, Poland

25CHI101: Statistical Learning and Medical Diagnostics
Room: C210
Organizer: Ngai Hang Chan
Chair: Jinfeng Xu

16:00 Boundary Detection and Image Segmentation via Local Discrepancy Scan Statistics

♦Richeng Hu, Ngai Hang Chan, Chung Wang Wong, Chun Yip Yau

The Chinese University of Hong Kong, City University of Hong Kong, The University of Hong Kong, The Chinese University of Hong Kong

16:25 AI-Powered Polyp Analysis in Colonoscopy: Improving Detection and Assessment

♦Jinfeng Xu

City University of Hong Kong

16:50 Predicting Future Change-points in Time Series

♦*Chun Yip Yau*

Chinese University of Hong Kong

17:15 Building an Artificial Intelligence-Based Infrastructure for Prospectively Validating Glaucoma Detection from 3D Optical Coherence Tomography Scans in real-world

♦*Anran Ran, Clement C. Tham, Carol Y. Cheung*

Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong

25CHI105: Statistical methods elevated by modern computation and massive data
Room: C301
Organizer: Yanyuan Ma
Chair: Jiwei Zhao

16:00 Likelihood-based Nonparametric Receiver Operating Characteristic Curve Analysis in the Presence of Imperfect Reference Standard

♦*Peijun Sang, Yifan Sun, Qinglong Tian, Pengfei Li*

University of Waterloo, University of Waterloo, University of Waterloo, University of Waterloo

16:25 Statistical Benefits when Incorporating LLM-Derived Predictions: Old Wine in a New Bottle?

♦*Jiwei Zhao*

University of Wisconsin - Madison

16:50 Robust Transfer Learning with Heterogeneous Data

Jing Wang, HaiYing Wang, ♦*Kun Chen*

University of Connecticut, University of Connecticut, University of Connecticut

25CHI106: Statistical Methods for Survival Data with Complex Censoring and Missing or Mismeasured Covariates
Room: C302
Organizer: Yanqing Sun
Chair: Yanqing Sun

16:00 Improving Estimation Efficiency for Case-cohort Studies with a Cure Fraction

♦*Qingning Zhou, Xu Cao*

University of North Carolina at Charlotte, University of California at Riverside

16:25 Evaluating predictive accuracy of prognostic models with interval-censored data

♦*Yang Qu, Yu Cheng*

Central South University, China, University of Pittsburgh, USA

16:50 A corrected smoothed score approach for semiparametric accelerated failure time model with error-contaminated covariates

♦*Xiao Song*

University of Georgia

17:15 A Flexible Copula Model for Bivariate Survival Data with Dependent Censoring

Reuben Adatorwovor, ♦*Yinghao Pan*

University of Kentucky, University of North Carolina at Charlotte

25CHI108: Statistical Methods in Medical Applications
Room: C303
Organizer: Jialiang Li
Chair: Jialiang Li

16:00 Heritability: a counterfactual perspective

Haochen Li, Jieru Shi, ♦*Hongyuan Cao, Qingyuan Zhao*

Florida State University, Cambridge University, Florida State University, Cambridge University

16:25 Estimation and Prediction of Time-in-range (TIR) with Inpatient Continuous Glucose Monitoring

Qi Yu, Guillermo Umpierrez, ♦*Limin Peng*

Emory University, Emory University, Emory University

16:50 Asymptotic distribution-free change-point detection for modern data based on a new ranking scheme

Doudou Zhou, ♦*Hao Chen*

National University of Singapore, University of California, Davis

17:15 Investigating, Interpreting, and Optimizing Wearable Device Usage in Diabetes Patients

♦*Jin Zhou, Bowen Zhang, Hua Zhou*

University of California, Los Angeles

University of California, Los Angeles

University of California, Los Angeles

25CHI112: Structured machine learning
Room: C304
Organizer: Junhui Wang
Chair: Ben Dai

16:00 Golden Ratio Weighting Prevents Model Collapse

Hengzhi He, ♦ *Shirong Xu, Guang Cheng*

University of California, Los Angeles

University of California, Los Angeles

University of California, Los Angeles

16:25 Two-way latent matching model for network analysis

♦*Ting Li, Jiangzhou Wang, Jianhua Guo*

The Hong Kong Polytechnic University, Shenzhen University, Beijing Gongshang University

16:50 Statistical Inference in Tensor Completion: Optimal Uncertainty Quantification and

Statistical-to-Computational Gaps

Wanteng Ma, ♦Dong Xia

University of Pennsylvania, Hong Kong University of Science and Technology

17:15 Decentralized learning of low-rank matrix

Zihao Song, Weihua Zhao, ♦Heng Lian

Nantong University, Nantong University, henglian@cityu.edu.hk

25CHI114: Transforming clinical trials with causal inference thinking and methodology
Room: C305
Organizer: Zhiwei Zhang
Chair: Min Zhang

16:00 A Connection Between Covariate Adjustment and Stratified Randomization in Randomized Clinical Trials

♦Zhiwei Zhang

Gilead Sciences

16:25 An adaptive design for optimizing treatment assignment in randomized clinical trials

♦Wei Zhang, Zhiwei Zhang, Aiyi Liu

Chinese Academy of Sciences, Gilead Sciences, National Institutes of Health

16:50 Incorporating external data for analyzing randomized clinical trials: A transfer learning approach

Yujia Gu, Hanzhong Liu, ♦Wei Ma

Renmin University of China, Tsinghua University, Renmin University of China

17:15 Joint Modeling of Longitudinal Biomarker and Survival Outcomes with the Presence of Competing Risk in Nested Case-Control Studies with Application to the TEDDY Microbiome Dataset

♦Jiyuan Hu

NYU Grossman School of Medicine

25CHI001: Advanced Experimental Design and Subsampling Approaches for Complex Data Analysis
Room: C405
Organizer: Mingyao Ai
Chair: Yaping Wang

16:00 Multi-resolution subsampling for linear classification with massive data

Haolin Chen, Holger Dette, ♦Jun Yu

Beijing Insititute of Technology, Ruhr-Universitat Bochum, Fakultat fur Mathematik, Beijing Insititute of Technology

16:25 A distance metric-based space-filling subsampling method for nonparametric models

Huaimin Diao, Dianpeng Wang, ♦Xu He

Shandong Technology and Business University, Beijing Institute of Technology, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

16:50 Stratum Order-of-Addition Designs

Liushan Zhou, Ze Liu, Min-Qian Liu, ♦Guanzhou Chen

Nankai University, Nankai University, Nankai University, Nankai University

17:15 Enhancing Sensitivity Analysis of Building Energy Performance through Batch-Sequential Maximum One-Factor-At-A-Time Designs

Qiang Zhao, Chunwei Zheng, Fasheng Sun, ♦Qian Xiao

Northeast Normal University, Nankai University, Northeast Normal University, Shanghai Jiao Tong University

25CHI002: Advanced Learning Methods for Complex Medical Data
Room: C306
Organizer: Hua Liang
Chair: Hua Liang

16:00 Parametric Modal Regression with Contaminated Covariates

Yanfei He, Jianhong Shi, ♦Weixing Song

Kansas State University

16:25 Unsupervised Domain Adaptation with Adaptive f-Divergence: Tighter Variational Representation and Generalization Bounds

♦Fode Zhang, Yifan Zhu, Zhe Cheng

Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, Southwestern University of Finance and Economics

16:50 Large-scale survival analysis with a cure fraction

Bo Han, ♦Xiaoguang Wang, Liuquan Sun

Yunnan University, Dalian University of Technology, Institute of Mathematics and Systems Science, Chinese Academy of Sciences

17:15 Survival Prediction in ALS Patients Using Deep Learning

♦Haiyan Su, George Li, Liuxia Wang

Montclair State University, Carnegie Melon University, Afinity

25CHI006: Advanced Statistical Methods for Spatial Transcriptomics
Room: C406
Organizer: Liping Zhu
Chair: Qing Cheng

16:00 Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics

♦Xiang Zhou

University of Michigan

16:25 Cross-technology and cross-resolution framework for spatial omics annotation with CAESAR

♦Jin Liu, Xiao Zhang, Wei Liu

The Chinese University of Hong Kong (Shenzhen), The Chinese University of Hong Kong (Shenzhen), Sichuan University

16:50 A de novo, spatially-aware and robust detection of phenotype-associated genes and tissue domains from multi-sample, multi-condition spatial transcriptomics

Wenlin Li, Yan Lu, Maocheng Zhu, Zhongkun Qu, Jin Liu, ♦Xiaobo Sun

School of Data Science, The Chinese University of Hong Kong-Shenzhen, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Data Science, The Chinese University of Hong Kong-Shenzhen, Department of Human Genetics, Emory University

17:15 Scaling up spatial transcriptomics for large-sized tissues

♦Mingyao Li

University of Pennsylvania

25CHI007: Advancements in High-Dimensional Statistical Methods and Applications
Room: C307
Organizer: Jinfeng Xu
Chair: Zhenggang Wang

16:00 Adaptive stratified sampling design in two-phase studies for average causal effect estimation

Min Zeng, Qiyu Wang, Zijian Sui, ♦Hong Zhang, Jinfeng Xu

City University of Hong Kong, University of Science and Technology of China, University of Science and Technology of China, University of Science and Technology of China, City University of Hong Kong

16:25 High-dimensional statistical methods for analyzing three-dimensional genomic data

♦Dechao Tian

Sun Yat-sen University

16:50 High-Dimensional Bias Propagation in Multi-Temporal Covariance Dynamics: A Random Matrix Theory Approach

♦Zhenggang Wang

Southeast University

17:15 On generalized transformation models

♦Zhezhen Jin

Columbia University

c. June 29th AM (8:30-10:10)

25CHI008: Advances in Causal Discovery for Omics Data
Room: C203
Organizer: Shanghong Xie
Chair: Haoran Xue

8:30 Robust Multi-ancestry PWAS Utilizing Bayesian Fine-mapping

Chengli Zhang, Chong Wu, ♦Haoran Xue

City University of Hong Kong, The University of Texas MD Anderson, City University of Hong Kong

8:55 A novel multivariable Mendelian randomization framework to disentangle highly correlated exposures with application to metabolomics

♦Lap Sum Chan, Mykhaylo Malakhov, Wei Pan

University of Minnesota, University of Minnesota, University of Minnesota

9:20 Leveraging Cross-population Fine-mapping to Strengthen cis-Mendelian Randomization in TWAS

♦Mingxuan Cai

City University of Hong Kong

9:45 Mitigating Subgroup Bias in Federated Selection of Sepsis Care Bundles

♦Yanyan Zhao, Peili Liu

Shandong University, Shandong University

25CHI011: Advances in Spatial Statistics with Random Field Modeling: Methods, Metrics, and Applications
Room: C204
Organizer: Juan Du
Chair: Juan Du

8:30 Estimation and model selection in general spatial dynamic panel data

Li Hou, Baisuo Jin, ♦Yuehua Wu

University of Science and Technology of China, University of Science and Technology of China, York University

8:55 Local Maxima of Discrete Gaussian Processes

♦Dan Cheng, John Ginos

Arizona State University, Arizona State University

9:20 A Bayesian nonstationary model for spatial binary data based on tree partition processes

♦Bohai Zhang, Furong Li, Jianxin Pan

Beijing Normal-Hong Kong Baptist University, Ocean University of China, Beijing Normal-Hong Kong Baptist University

9:45 Time-varying vector random fields on the arccos-quasi-quadratic metric space

Juan Du, ♦Chunsheng Ma

Kansas State University, Wichita State University

25CHI014: Advances in Statistical Methods and Applications
Room: C205
Organizer: Lizhe Sun
Chair: Lizhe Sun

8:30 Stochastic feature selection with annealing and its applications to streaming data

♦*Lizhe Sun, Adrian Barbu*

Shanxi University of Finance and Economics, Florida State University

8:55 A comparison of two models for detecting inconsistency in network meta-analysis

♦*Lu Qin, Shishun Zhao, Wenlai Guo, Tiejun Tong, Ke Yang*

Center for Applied Statistical Research and College of Mathematics, Jilin University, Changchun, China, Center for Applied Statistical Research and College of Mathematics, Jilin University, Changchun, China, Department of Hand Surgery, the Second Hospital of Jilin University, Changchun, China, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China, Department of Statistics and Data Science, Beijing University of Technology, Beijing, China

9:20 Distance-based Clustering of Functional Data with Derivative Principal Component Analysis

♦*Ping Yu, Gongming Shi, Chunjie Wang, Xinyuan Song*

Shanxi Normal University, Capital University of Economics and Business, Changchun University of Technology, The Chinese University of Hong Kong

9:45 Testing for the equality of distributions in high dimension

♦*Xu Li, Gongming Shi, Baoxue Zhang*

Shanxi Normal University, Capital University of Economics and Business, Capital University of Economics and Business

25CHI017: Advances in Statistical Modeling: Variable Selection, Dependence, and Nonparametric Methods
Room: C207
Organizer: Hongmei Jiang
Chair: Hongmei Jiang

8:30 BELIEF in Dependence

Benjamin Brown, ♦Kai Zhang, Xiao-Li Meng,

UNC Chapel Hill, UNC Chapel Hill, Harvard University

8:55 Variable selection for partially linear models and partially global Fréchet regression

♦*Yichao Wu*

University of Illinois Chicago

9:20 A new approach to select linear and nonparametric predictors simultaneously for generalized partially linear models

Youhan Lu, ♦Juan Hu, Yichao Wu

University of Illinois Chicago, DePaul University, University of Illinois Chicago

9:45 Quantile estimation for nonparametric regression models with autoregressive and moving average errors

♦*Qi Zheng, Yunwei Cui*

University of Louisville, Townson University

25CHI021: Causal inference and decision-making
Room: C208
Organizer: Yifan Cui
Chair: Yuanshan Gao

8:30 Proximal Inference on Population Intervention Indirect Effect

♦*Yang Bai, Yifan Cui, Baoluo Sun*

National University of Singapore, Zhejiang University, National University of Singapore

8:55 Learning Robust Treatment Rules for Censored Data

Yifan Cui, Junyi Liu, ♦Tao Shen, Zhengling Qi, Xi Chen

Zhejiang University, Tsinghua University, National University of Singapore, George Washington University, New York University

9:20 Causal mediation analysis of data fusion with application to bridging risk and relative efficacy of vaccines

♦*Pan Zhao, Oliver Dukes, Bo Zhang*

University of Cambridge, Ghent University, Fred Hutchinson Cancer Center

25CHI027: Frontier Statistical Methods for Single-cell RNA Sequencing Data
Room: C209
Organizer: Xiaodan Fan
Chair: Xiaodan Fan

8:30 scTEL: Protein Expression Prediction in Single-cell Analysis Using Transformer

♦*Chaojie Wang*

Jiangsu University

8:55 Large-scale imputation of spliced and unspliced RNA counts for Cell Lineage analysis

♦*Shanjun Mao*

Hunan University

9:20 Temporal mapping and clonal differentiation modelling from time-series single-cell RNA-seq data

Yijun Liu, Mingze Gao, ♦Yuanhua Huang

University of Hong Kong, University of Hong Kong, University of Hong Kong

25CHI030: Innovations and Partnerships in Data-Rich Environments: Emerging Advances in Matrix and Tensor

Modeling
Room: C404
Organizer: Jiangyan Wang
Chair: Jiangyan Wang

8:30 Shape Mediation Analysis in Alzheimer's Disease Studies

Xingcai Zhou, Miyeon Yeon, ♦Jiangyan Wang, Shengxian Ding, Kaizhou Lei, Yanyong Zhao, Rongjie Liu, Chao Huang

Nanjing Audit University

8:55 Matrix-factor-augmented regression

♦Xiong Cai, Xinbing Kong, Xinlei Wu, Peng Zhao

Nanjing Audit University, Southeast University, Nanjing Audit University, Jiangsu Normal University

9:20 Matrix-quantile factor prediction for generalized matrix-variate regression

♦Yongxin Liu

Nanjing Audit University

9:45 A Model-Based Monitoring Framework for Tensor Count Data in Passenger Flow Surveillance

♦Yifan Li

Nanjing Audit University

25CHI041: Integrate Statistics into Deep Learning for Digital Image Processing and Analysis
Room: C210
Organizer: Weihong Guo
Chair: Junying Meng

8:30 Solving Unbalanced Optimal Transport on Point Cloud by Tangent Radial Basis Function Method

♦Jiangong Pan

Tsinghua University

8:55 Learnable Mixture Distribution Prior for Deep Learning based Image Processing

♦Jun Liu

Beijing Normal University

9:20 Learnable Nonlocal Self-similarity of Deep Features for Image Denoising

♦Junying Meng, Faqiang Wang, Jun Liu

Shanxi University, Beijing Normal University, Beijing Normal University

25CHI044: Machie learning for data assimilation
Room: C301
Organizer: Yuling Jiao
Chair: Shuyi Zhang

8:30 High-dimensional Ensemble Kalman Filter with Localization, Inflation and Iterative Updates

♦Hao-Xuan Sun, Shouxia Wang, Xiaogu Zheng, Song Xi Chen

Peking University, Beijing, China, Shanghai University of Finance and Economics, Shanghai, China, International Global Change Institute, Hamilton, New Zealand, Tsinghua University, Beijing, China

8:55 Generative Assimilation Forecasting

♦Baoxiang Pan

Institute of Atmospheric Physics, Chinese Academy of Science

9:20 Nonlinear assimilation with score-based sequential Langevin sampling

♦Cheng Yuan

Huazhong Normal University

25CHI046: Modeling and inference for distributions and high dimensional data
Room: C302
Organizer: Ming-Yen Cheng
Chair: Jialiang Li

8:30 Penalized weighted generalized estimation equations for high-dimensional longitudinal data with informative cluster size

♦Haofeng Wang

Hong Kong Baptist university

8:55 Two-sample tests for equal distributions in separable metric spaces: a unified semimetric-based approach

♦Jin-Ting Zhang, Meichen Qian, Tianming Zhu

National University of Singapore, National University of Singapore, Nanyang Technological University

9:20 Generalized Median of Means Principle for Bayesian Inference

Stanislav Minsker, ♦Shunan Yao

University of Southern California, Hong Kong Baptist University

9:45 Oracle-efficient estimation and trend inference in non-stationary time series with trend and heteroscedastic ARMA error

♦Chen Zhong

Fuzhou University

25CHI059: Novel Machine Learning Methods for Disease Progression and Precision Medicine
Room: C304
Organizer: Shanghong Xie
Chair: Huichen Zhu

8:30 Dynamic System for Modeling Latent Disease Progression and Treatment Effect

♦Zexi Cai, Yuanjia Wang, Shanghong Xie

Columbia University, Columbia University, University of South Carolina

8:55 LATENT GAUSSIAN PROCESS JOINT MODEL FOR INTEGRATIVE ANALYSIS OF MULTIMODAL BIOMARKERS AND INITIATION

OF MEDICATION OF PARKINSON'S DISEASE

Junxuan Chen, Xiangnan Feng, ♦Kai Kang

Sun Yat-sen University, Fudan University, Sun Yat-sen University

9:20 Heterogeneous Quantile Treatment Effect Estimation with High-Dimensional Confounding

♦Huichen Zhu

The Chinese University of Hong Kong

25CHI061: Observational data analysis with complex study designs
Room: C305
Organizer: Andy Ni
Chair: Yuzi Zhang

8:30 Stochastic Explicit Calibration Algorithm for Survival Models

♦Jeongho Park

Yonsei University

8:55 Leveraging LLM-Derived Gene Embeddings for Gene-Expression Analysis

♦Jun Li

University of Notre Dame

9:20 Exponential Power Mixture of Experts Model: Estimation, Clustering, and Variable Selection

♦Zhenghui Feng, Xuefei Qi, Heng Peng, Xingbai Xu, Jie Xue

Harbin Institute of Technology, Xiamen University, HongKong Baptist University, Xiamen University, Xiamen University

9:45 Kernel Ridge Regression with Predicted Feature Inputs and Applications to Factor-Based Nonparametric Regression

Xin Bing, Xin He, ♦Chao Wang

University of Toronto, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

25CHI063: Random matrices and high-dimensional statistics
Room: C306
Organizer: Jeff Yao
Chair: Jeff Yao

8:30 Eigenvalues of large dimensional information plus noise type matrices

♦Huanchao Zhou, Zhidong Bai, Jiang Hu, Jack Silverstein

School of Mathematics and Computational Science, Xiangtan University, , School of Mathematics and Statistics, Northeast Normal University, School of Mathematics and Statistics, Northeast Normal University, Department of Mathematics, North Carolina State University

8:55 On spiked eigenvalues of a renormalized sample covariance matrix from multi-population

Weiming Li, Zeng Li, ♦Junpeng Zhu

Shanghai University of Finance and Economics, Southern University of Science and Technology, Southern University of Science and Technology

9:20 Limiting spectral distribution for a cross data matrix-based matrix

♦Shao-Hsuan Wang

National Central University

25CHI069: Recent Advances in Microbiome Data Analysis
Room: C405
Organizer: Tao Wang
Chair: Tiantian Liu

8:30 Knockoff-based high-dimensional mediator identification and its application in microbiome research

♦Tiantian Liu, Dong Xu

China Pharmaceutical University, Shanghai Jiaotong University

8:55 gmmcoda: Graphical model for the mixture of compositional data and absolute abundance data with applications to microbiome studies

Shen Zhang, ♦Huaying Fang, Tao Hu

Capital Normal University, Capital Normal University, Capital Normal University

9:20 Integrative analysis of microbial 16S gene and shotgun metagenomic sequencing data improves statistical efficiency in testing differential abundance

♦Yicong Mao, Ye Yue, Timothy Read, Veronika Fedirko, Glen Satten, Xuan Chen, Xiang Zhan, Yi-Juan Hu

Department of Biostatistics, Peking University, Department of Biostatistics and Bioinformatics, Emory University, , Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Department of Epidemiology, University of Texas MD Anderson Cancer Center & Department of Epidemiology, Emory University, Department of Gynecology and Obstetrics, Emory University School of Medicine, College of Economics and Management, Huazhong Agricultural University, School of Statistics and Data Science, Southeast University, Department of Biostatistics, School of Public Health, Peking University & Beijing International Center for Mathematical Research, Peking University & Center for Statistical Science, Peking University,

9:45 Differential abundance analysis of sequence count data

Guanxun Li, Xianyang Zhang, ♦Huijuan Zhou

Beijing Normal University at Zhuhai, Texas A&M Univerisity, Shanghai University of Finance and Economics

25CHI073: Recent Advances in Single-cell Data Analysis
Room: C307
Organizer: Tao Wang
Chair: Tao Wang

8:30 GraphPCA: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data

Jiyuan Yang, Lu Wang, Lin Liu, ♦Xiaoqi Zheng

Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University School of Medicine, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai Jiao Tong University School of Medicine

8:55 Sparse representation learning for scalable single-cell RNA sequencing data analysis

Kai Zhao, Hon-Cheong So, ♦Zhixiang Lin

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

9:20 Decoding Gene Functions: Exploring their Significance in Biological Context

♦Ying Zhu

Fudan University

9:45 Generative Modeling of Single-cell Dynamics with Deep Diffusion Schrödinger Bridge Model

♦Jingsi Ming

East China Normal University

d. June 29th AM (10:30-12:10)

25CHI075: Recent Advances in Statistical Machine Learning: Theory and Algorithms
Room: C203
Organizer: Yunwen Lei
Chair: Yunwen Lei

10:30 Word-Level Maximum Mean Discrepancy Regularization for Word Embedding

Youqian Gao, ♦Ben Dai

The Chinese University of Hong Kong, The Chinese University of Hong Kong

10:55 Functional data analysis via neural networks

♦Jun Fan

Hong Kong Baptist University

11:20 Understanding token selection in the self-attention mechanism

Zihao Li, ♦Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason Klusowkski, Jianqing Fan, Mengdi Wang

Princeton University, The University of Hong Kong, Princeton University, Princeton University, Northwestern University, Princeton University, Princeton University, Princeton University

11:45 Why Does Differential Privacy Noise Have Limited Impact When Fine-Tuning Large Language Models?

♦Chendi Wang

Xiamen Univeristy

25CHI076: Recent Advances on the Analysis of Failure Time Data
Room: C204
Organizer: Jianguo Sun
Chair: Jianguo Sun

10:30 Estimation and Variable Selection for Interval-Censored Failure Time Data with Random Change Point and Application to Breast Cancer Study

♦Mingyue Du, Yichen Lou, Jianguo Sun

Jilin University, The Chinese University of Hong Kong, University of Missouri

10:55 Linearized maximum rank correlation estimation of doubly truncated data

♦Peijie Wang, Qihao Wang, Jianguo Sun

Jilin University, Jilin University, University of Missouri

11:20 Goodness of fit test for bivariate interval-censored survival data

Bernard Rosner, Camden Bay, Robert Glynn, Gui-Shuan Ying, Maureen Maguire, ♦Mei-Ling Ting Lee

Harvard Medical School, Brigham and Women's Hospital, Harvard Medical School, University of Pennsylvania, University of Pennsylvania, University of Maryland

11:45 Bayesian estimation of partial functional Tobit censored quantile regression model

♦Chunjie Wang, Zhexin Lu, Chuchu Wang, Xinyuan Song

Cahngchun University of Technology, Cahngchun University of Technology, The Chinese University of Hong Kong, The Chinese University of Hong Kong

25CHI078: Recent development in statistical methods for related regression models and applications
Room: C205
Organizer: Zhiqiang Cao
Chair: Jie He

10:30 Sequential quantile regression for streaming data by least squares

♦Ye Fan, Nan Lin

Capital University of Economics and Business, Washington University in St. Louis

10:55 Assessing mediation in cross-sectional stepped wedge cluster randomized trials

♦Zhiqiang Cao, Fan Li

Shenzhen Technology University, Yale University

11:20 Nonparametric Sensitivity Analysis for Unobserved Confounding with Survival Outcomes

♦Rui Hu, Ted Westling

Shenzhen Technology University, University of Massachusetts Amherst

11:45 Matrix Autoregressive Time Series with Reduced-Rank and Sparse Structural Constraints

♦Xiaohang Wang, Ling Xin, Philip L.H. Yu

The Education University of Hong Kong, Beijing Normal-Hong Kong Baptist University, The Education University of Hong Kong

25CHI079: Recent developments about high-dimensional inference
Room: C207
Organizer: Xu Guo
Chair: Long Feng

10:30 Targeted Inference for High-Dimensional Quantile Regression Models

♦Xuejun Jiang, Yakun Liang

Department of Statistics and Data Science, Southern University of Science and Technology, Department of Statistics and Data Science, Southern University of Science and Technology

10:55 Robust Mutual Fund Selection with False Discovery Rate Control

Hongfei Wang, Ping Zhao, ♦Long Feng, Zhaojun Wang

Nanjing Audit University, Nankai University, Nankai University, Nankai University

11:20 Homogeneity tests of high-dimensional covariance matrices with applications to change-points detection

Jiayu Lai, Xiaoyi Wang, ♦Le Zhou, Shurong Zheng

Northeast Normal University, Beijing Normal University, Hong Kong Baptist University, Northeast Normal University

11:45 Testing the general linear hypothesis in high-dimensional heteroscedastic factor model via random integration

♦Mingxiang Cao

Anhui Normal University

25CHI081: Recent developments in causal inference and survival analysis
Room: C208
Organizer: Yifan Cui
Chair: Yuanshan Gao

10:30 SEMIPARAMETRIC CURE REGRESSION MODELS WITH INFORMATIVE CASE K INTERVAL-CENSORED FAILURE TIME DATA

♦Yichen Lou, Jianguo Sun, Peijie Wang

The Chinese University of Hong Kong, University of Missouri, Jilin University

10:55 Significance test for semiparametric conditional average treatment effects and other structural functions

♦Niwen Zhou, Xu Guo, Lixing Zhu

Beijing Normal University, Beijing Normal University, Beijing Normal University

11:20 Causal inference for time-to-event data with a cured subpopulation

♦Yi Wang, Yuhao Deng, Xiaohua Zhou

Shanghai University of International Business and Economics, Peking University, Peking University

11:45 Recent developments in causal inference and survival analysis

Chang Wang, ♦Baihua He, Shishun Zhao, Jianguo Sun, Xinyu Zhang

University of Science and Technology of China

25CHI082: Recent Developments in Covariate Adjustment for Randomized Clinical Trials
Room: C209
Organizer: Xin Zhang
Chair: Xin Zhang

10:30 Bias Reduction in G-computation for Covariate Adjustment in Randomized Clinical Trials

♦Xin Zhang, Lin Liu, Haitao Chu

Pfizer Inc, Shanghai Jiao Tong University, Pfizer Inc and University of Minnesota

10:55 Robust and Efficient Statistical Inference Under Covariate-Adaptive Randomization

♦Fuyi Tu

School of Science, Chongqing University of Posts and Telecommunications

11:20 Covariate Adjustment in Randomized Experiments Motivated by Higher-Order Influence Functions

♦Lin Liu

Shanghai Jiao Tong University

11:45 Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates

Xin Lu, Fan Yang, ♦Yuhao Wang

Tsinghua University, Tsinghua University, Tsinghua University

25CHI083: Recent Developments in High-dimensional Data Analysis
Room: C210
Organizer: Xingqiu Zhao
Chair: Xingqiu Zhao

10:30 Integrative Analysis of High-dimensional RCT and RWD Subject to Censoring and Hidden

Confounding

♦*Xin Ye, Shu Yang, Xiaofei Wang, Yanyan Liu*

School of Statistics and Mathematics, Guangdong University of Finance and Economics, Department of Statistics, North Carolina State University, Department of Biostatistics and Bioinformatics, Duke University, School of Mathematics and Statistics, Wuhan University

10:55 Kernel Variable Importance Measure with Applications

♦*Bingyao Huang, Guanghui Cheng, Yanyan Liu, Liuhua Peng*

Guangdong University of Technology, Guangzhou University, Wuhan University, The University of Melbourne

11:20 Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

♦*Xiangbin Hu*

Beijing Institute of Technology

11:45 Non-parametric inference based on reliability life-test of non-identical coherent systems

♦*Xiaojun Zhu*

Xi'an Jiaotong-Liverpool University

25CHI084: Recent developments in reinforcement learning and mobile health
Room: C301
Organizer: Yifan Cui
Chair: Tao Shen

10:30 Factorial Causal Excursion Effects: Modeling Time-Varying Effects of Multi-Component Mobile Health Interventions in Micro-Randomized Trials

♦*Xueqing Liu, Weihao Li, Bibhas Chakraborty*

Duke-NUS Medical School, National University of Singapore, Duke-NUS Medical School

10:55 Minimax Regret Learning for Data with Heterogeneous Sub-populations

♦*Weibin Mo, Weijing Tang, Songkai Xue, Yufeng Liu, Ji Zhu*

Purdue University, Carnegie Mellon University, University of Michigan, University of North Carolina, Chapel Hill, University of Michigan

11:20 A Synergetic Random Forest Framework for Policy Evaluation

Rui Qiu, Zexuan Zhang, Zhou Yu, ♦*Ruoqing Zhu*

Beiing University, University of Illinois Urbana Champaign, East China Normal University, University of Illinois Urbana Champaign

11:45 Online statistical inference for robust policy evaluation in reinforcement learning

Weidong Liu, Jiyuan Tu, Xi Chen, ♦*Yichen Zhang*

SJTU, SUFE, New York University, Purdue University

25CHI088: Robust Analysis for Treatment Decision and Risk Prediction under Complex Data Settings
Room: C302
Organizer: Donglin Zeng
Chair: Yu Gu

10:30 Semiparametric Regression Analysis for Interval-Censored Outcome Subject to Misdiagnosis

♦*Yuhao Deng, Donglin Zeng, Yuanjia Wang*

University of Michigan, University of Michigan, Columbia University

10:55 INFERENCE FOR HIGH DIMENSIONAL PROPORTIONAL HAZARDS MODEL WITH STREAMING SURVIVAL DATA

♦*Dongxiao Han*

School of Statistics and Data Science, Nankai University

11:20 A Robust Covariate-Balancing Method for Estimating Individualized Treatment with Censored Data

Rujia Zheng, ♦*Wensheng Zhu, Xiaofan Guo*

Northeast Normal University, Yunnan University, The First Hospital of China Medical University

11:45 Learning Optimal Early Decision Treatment Rules with Multi-domain Intermediate Outcomes

♦*Yuanjia Wang*

Columbia University

25CHI093: Some recent developments about big data analysis
Room: C304
Organizer: Xu Guo
Chair: Lei Wang

10:30 Communication-Efficient and Distributed-Oracle Estimation for High-Dimensional Quantile Regression

♦*Songshan Yang, Yifan Gu, Hanfang Yang, Xuming He*

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Center for Applied Statistics and School of Statistics, Renmin University of China, Center for Applied Statistics and School of Statistics, Renmin University of China, Department of Statistics and Data Science, Washington University in St. Louis

10:55 Optimal subsampling for high-dimensional partially linear models via machine learning methods

Yujing Shao, ♦*Lei Wang, Heng Lian, Haiying Wang*

Nankai University

11:20 Least Squares and Hypothesis Testing based Transfer Learning for High-Dimensional Quantile Regression

♦*Kangning Wang, Xiaotong Zhu*

Shandong Technology and Business University, Shandong Technology and Business University

25CHI096: Statistical Advances in Large Language Models and Network Analysis
Room: C305
Organizer: Will Wei Sun
Chair: Will Wei Sun

10:30 Learning nonparametric graphical model on heterogeneous network-linked data

♦Junhui Wang

Chinese University of Hong Kong

10:55 A Statistical Hypothesis Testing Framework for Data Misappropriation Detection in Large Language Models

Yinpeng Cai, Lexin Li, ♦Linjun Zhang

Peking University, UC Berkeley, Rutgers University

11:20 Robust Reinforcement Learning from Human Feedback for Large Language Models Fine-Tuning

Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, ♦Chengchun Shi

LSE, Tsinghua University, LSE, Oxford, London School of Economics and Political Science

11:45 A Statistical Take on Watermarks for Large Language Models: Theory, Applications, and Future Opportunities

♦Weijie Su

University of Pennsylvania

25CHI098: Statistical analysis of complex survival data
Room: C306
Organizer: Chunjie Wang
Chair: Shuying Wang

10:30 Semiparametric Analysis of Additive–Multiplicative Hazards Model with Interval-Censored Data and Panel Count Data

Tong Wang, Yang Li, Jianguo Sun, ♦Shuying Wang

Changchun University of Technology, Changchun University of Technology, University of Missouri, Changchun University of Technology

10:55 Group penalized doubly nonparametric probit model with interval censored survival data and application to pharyngeal disease

♦Bo Zhao, Chunjie Wang, Shuying Wang, Dan Yu

Changchun University of Technology, Changchun University of Technology, Changchun University of Technology, Department of Otolaryngology Head and Neck Surgery the Second Hospital, Jilin University

11:20 Distributed Least Product Relative Error estimation for semi-parametric multiplicative regression with massive data

♦Xiaohui Yuan

Changchun University

11:45 Bayesian empirical likelihood for accelerated failure time model with covariates missing at random

♦Xinrui Liu

Changchun University of Technology

25CHI111: Statistics for Emerging Trends in Machine Learning
Room: C307
Organizer: Yafei Wang
Chair: Yafei Wang

10:30 Advancing Fairness in Healthcare: A Universal Framework for Optimal Treatment Effect Estimation with Censored Data

Hongni Wang, ♦Junxi Zhang, Na Li, Linglong Kong, Bei Jiang, Xiaodong Yan

Shandong University of Finance and Economics, Concordia University, Shandong University of Finance and Economics, University of Alberta, University of Alberta, Xi'an Jiaotong University

10:55 An adaptive model checking test for the functional linear model

Enze Shi, Yi Liu, Ke Sun, ♦Lingzhu Li, Linglong Kong

University of Alberta, University of Alberta, University of Alberta, Beijing University of Technology, University of Alberta

11:20 Some Recent Optimal Exact Confidence Intervals in Contingency Tables

♦Weizhen Wang

Beijing University of Technology

11:45 Online model averaging prediction

♦Jun Liao

Renmin University of China

25CHI024: Design and analysis of clinical studies
Room: C404
Organizer: Zhezhen Jin
Chair: Zexi Cai

10:30 A Generalized Outcome-Adaptive Sequential Multiple Assignment Randomized Trial Design

Xue Yang, ♦Yu Cheng, Peter Thall, Wabdus Wahed

University of Pittsburgh, University of Pittsburgh, University of Texas MD Anderson Cancer Center, University of Rochester

10:55 State-dependent sampling designs for prevalent cohort studies

♦Leilei Zeng
University of Waterloo

11:20 Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

Xiangbin Xiangbin, ♦Wen Su, Zhisheng Ye, Xingqiu Zhao

The Hong Kong Polytechnic University, City

University of Hong Kong, National University of Singapore, The Hong Kong Polytechnic University

25CHI065: Recent Advances in Complex Data

Room: C405
Organizer: Yang Zhou
Chair: Yang Zhou

10:30 Ordinary Differential Equation Models for a Collection of Discretized Functions

♦Lingxuan Shao, Fang Yao

Fudan University, Beijing University

10:55 Regularized reduced-rank regression for structured output prediction

Heng Chen, Di-Rong Chen, ♦Kun Cheng, Yang Zhou

Capital University of Economics and Business Beihang University Beijing Jiaotong University Beijing Normal University

11:20 Two-Sample Distribution Tests in High Dimensions via Max-Sliced Wasserstein Distance and Bootstrapping

♦Xiaoyu Hu, Zhenhua Lin

Xi'an Jiaotong University National University of Singapore

11:45 Change-Points Detection and Support Recovery for Spatiotemporal Functional Data

♦Decai Liang

Nankai University

e. June 29th PM (14:00-15:40)

25CHI003: Advanced Statistical and Computational Methods for Microbiome and Metagenomics Data Analysis

Room: C203
Organizer: Yanan Zhao
Chair: Jiyuan Hu

14:00 Integrating functional and taxonomic profiles for microbiome biomarker identification and disease prediction

♦Chan Wang, Huilin Li

NYU School of Medicine, NYU School of Medicine

14:25 TEMPTED: time-informed dimensionality reduction for longitudinal microbiome studies

♦Anru Zhang, Rungang Han, Yanan Zhao

Duke University, Duke University, Duke University

14:50 Joint modeling of longitudinal and time-to-event outcomes under nested case-control sampling with application to TEDDY biomarker study

♦Yanan Zhao, Jiyuan Hu

Department of Population and Health, NYU Grossman School of Medicine, Department of Population and Health, NYU Grossman School of Medicine

25CHI010: Advances in Modern Statistical Methodologies: Robust Estimation, High-Dimensional Inference, and Innovative Biomedical Applications

Room: C204
Organizer: Xiaoya Xu
Chair: Xiaoya Xu

14:00 Robust and Optimal Tensor Estimation via Robust Gradient Descent

♦Xiaoyu Zhang

Tongji University

14:25 Comparing MCP-MOD and Ordinal Linear Contrast Test in Dose Finding Clinical Trials: A Thorough Examination

Yaohua Zhang, ♦Ning Li, Naitee Ting

Boston University, Sanofi, StatsVita

14:50 Bayesian Design for Bridging Studies: Methods and Applications

♦Lichang Chen

Akeso Biopharma Inc.

15:15 Bootstrap inference for high dimensional nonconvex penalised regression and post-selection least squares

♦Xiaoya Xu, Stephen Lee

Shenzhen Polytechnic University, The University of Hong Kong

25CHI018: Advancing Multi-platform and Multi-modal Omics Harmonization

Room: C404
Organizer: Qian Li
Chair: Wei Liu

14:00 Spotiphy enables single-cell spatial whole transcriptomics across an entire section

♦Jiyuan Yang

Department of Computational Biology, St. Jude Children's Research Hospital

14:25 MODE: high-resolution digital dissociation with deep multimodal autoencoder

Jiao Sun, Tong Lin, Kyle Smith, Wei Zhang, Paul Northcott, ♦Qian Li

St. Jude Children's Research Hospital St. Jude Children's Research Hospital St. Jude Children's Research Hospital University of Central Florida St. Jude Children's Research Hospital St. Jude Children's Research Hospital

14:50 Inferring cell type-specific co-methylation networks from single-cell DNA methylation data

♦Jiebiao Wang

University of Pittsburgh

15:15 A deconvolution framework that uses single-cell sequencing plus a small benchmark data set for accurate analysis of cell type ratios in complex tissue samples

Shuai Guo, ♦Xiaoqian Liu, Xuesen Cheng, Rui Chen, Wenyi Wang

The University of Texas MD Anderson Cancer Center University of California, Riverside Baylor College of Medicine Baylor College of Medicine The University of Texas MD Anderson Cancer Center

25CHI019: Advancing Multi-Regional Clinical Trials: Methodology and Application Considerations for Ensuring Global Representation, Regulatory Harmonization, and Ethical Integrity

Room: C405
Organizer: Menggang Yu
Chair: Menggang Yu

14:00 Consistency Assessment of Treatment Effect in Multi-Regional Clinical Trials in the Presence of Treatment Effect Heterogeneity

♦Menggang Yu, Kunhai Qing

University of Michigan, East China Normal University

14:25 Considerations of China joining MRCTs - A case study in Central Nervous System (CNS) Therapeutic Area

♦Heli Gao, Hui Wang

Boehringer Ingelheim (China) Investment Co., Ltd. Boehringer Ingelheim (China) Investment Co., Ltd.

14:50 Regional consistency evaluation and sample size calculation under two MRCTs

Kunhai Qing, Xinru Ren, Shuping Jiang, Ping Yang, Menggang Yu, ♦Jin Xu

East China Normal University East China Normal University MSD China MSD China University of Michigan School of Public Health East China Normal University

15:15 Practical Considerations on Bayesian Hierarchical Models to Support Regional Effect Evaluation in Multi-Regional Clinical Trials

♦Rexin Lin

Novartis China

25CHI025: Dimension Reduction Methods

Room: C205
Organizer: Xin (Henry) Zhang
Chair: Ning Wang

14:00 Fast fitting of Gaussian mixture model via dimension reduction

♦Yin Jin, Wei Luo

Zhejiang University Zhejiang University

14:25 Robust Sliced Inverse Regression: Optimal Estimation for Heavy-Tailed Data in High Dimensions

♦Jing Zeng, Keqian Min, Qing Mai

University of Science and Technology of China IBM Florida State University

14:50 A Reduced-Rank Factor Model for Panel Data

Mingke Zhang, ♦Yingcun Xia

National University of Singapore National University of Singapore

25CHI028: High dimensional statistics inference

Room: C207
Organizer: Guangming Pan
Chair: Guangming Pan

14:00 Nonlinear Principal Component Analysis with Random Bernoulli Features for Process Monitoring

Ke Chen, ♦Dandan Jiang

Xi'an Jiaotong University Xi'an Jiaotong University

14:25 Portmanteau statistics for high-dimensional vector moving average processes

♦Chi Yao, Zeqin Lin, Xuejun Wang, Yiming Liu, Guangming Pan

Nanyang Technological University Nanyang Technological University Anhui University Jinan University Nanyang Technological University

14:50 Inference in Randomized Least Squares and PCA via Normality of Quadratic Forms

Leda Wang, ♦Zhixiang Zhang, Edgar Dobriban

Yale University University of Macau University of Pennsylvania

15:15 Distribution-Free and Model-Agnostic Changepoint Detection with Finite-Sample Guarantees

Xiaolong Cui, Haoyu Geng, ♦Guanghui Wang, Zhaojun Wang, Changliang Zou

Nankai University Nankai University Nankai University Nankai University Nankai University

25CHI029: High dimensional statistics inference (II)

Room: C208
Organizer: Guangming Pan
Chair: Bo Zhang

14:00 Huber Principal Component Analysis for large-dimensional factor models

Yong He, Lingxiao Li, ♦Dong Liu, Wen-Xin Zhou

Institute for Financial Studies, Shandong University Department of Data Science and Artificial Intelligence, Hong Kong Polytechnic University School of Physical and Mathematical Sciences, Nanyang Technological University College of Business Administration, University of Illinois at

Chicago

14:25 Revisiting the Spanning Puzzle of Bond Returns: A Unified Econometric Inference based on Predictive Quantile Regression

♦*Xiaohui Liu, Xinyi Wei*, Wei Long

Jiangxi University of Finance and Economics, Jiangxi University of Finance and Economics, Department of Economics, Tulane University

14:50 Identify the source of spikes: factor or mixture?

Yiming Liu

Jinan University

15:15 Identifying the structure of high-dimensional time series via eigen-analysis

♦*Bo Zhang, Jiti Gao, Guangming Pan, Yanrong Yang*

University of Science and Technology of China Monash University Nanyang Technological University Australian National University

25CHI031: Innovations in Causal Inference and Statistical Methods for Complex Data Structures

Room: C209
Organizer: Yidong Zhou
Chair: Doudou Zhou

14:00 The synthetic instrument: From sparse association to sparse causation

♦*Dingke Tang, Dehan Kong, Linbo Wang*

Washington University in St. Louis University of Toronto University of Toronto

14:25 Geodesic Causal Inference

Daisuke Kurisu, ♦*Yidong Zhou, Taisuke Otsu, Hans-Georg Müller*

Center for Spatial Information Science, The University of Tokyo Department of Statistics, University of California, Davis Department of Economics, London School of Economics Department of Statistics, University of California, Davis

14:50 Generalized Independence Test for Modern Data

♦*Mingshuo Liu, Doudou Zhou, Hao Chen*

UCDavis, NUS, UCDavis

15:15 Modeling Interval-Censored Outcome Data with a Potentially Interval-Censored Covariate

♦*Dongdong Li, Yue Song, Wenbin Lu, Huldrych Gunthard, Roger Kouyos, Rui Wang*

Harvard Medical School

25CHI034: Innovations in Nonparametric and Functional Data Analysis

Room: C210
Organizer: Xingche Guo
Chair: Xingche Guo

14:00 Semiparametric mixture regression for asynchronous longitudinal data using multivariate functional principal component analysis

♦*Yehua Li, Ruihan Lu, Weixin Yao*

University of California, Riverside US FDA University of California, Riverside

14:25 Estimation and Inference for Nonparametric Expected Shortfall Regression over RKHS

Myeonghun Yu, ♦*Yue Wang, Siyu Xie, Kean Ming Tan, Wenxin Zhou*

University of Michigan University of Science and Technology of China Northwestern University University of Michigan University of Illinois at Chicago

14:50 Variable Selection and Minimax Prediction in High-dimensional Functional Linear Models

♦*Xingche Guo, Yehua Li, Tailen Hsing*

University of Connecticut University of California, Riverside University of Michigan

15:15 Bias-Correction and Test for Mark-Point Dependence with Replicated Marked Point Processes

Ganggang Xu, Jingfei Zhang, Yehua Li, ♦*Yongtao Guan*

University of Miami Emory University UC Riverside Chinese University of Hong Kong, Shenzhen

25CHI043: Lifetime Data Analysis

Room: C301
Organizer: Mei-Ling Ting Lee
Chair: Mei-Ling Ting Lee

14:00 Semiparametric analysis of recurrent disease status data collected at intermittent follow-up

Yong-Chen Huang, ♦*Shu-Hui Chang*

National Taiwan University, National Taiwan University

14:25 A Varying-coefficient Additive Hazard Model for Recurrent Events Data

♦*Jialiang Li*

NUS

14:50 Deep Nonparametric Inference for Censored Data

Wen Su, Qiang Wu, Kin-Yat Liu, Guosheng Yin, Jian Huang, ♦*Xingqiu Zhao*

City University of Hong Kong, The Hong Kong Polytechnic University, The Chinese University of Hong Kong, The University of Hong Kong, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

15:15 Time-Adapted Exponential Models for Recurrent Covariates in Survival Analysis

Guangyu Yang, Boxian Wei, ♦*Min Zhang*

Institute of Statistics and Big Data, Renmin University of China Vanke School of Public Health, Tsinghua University Vanke School of Public Health, Tsinghua University

25CHI045: Model fairness and challenges and development of statistical models

Room: C302
Organizer: Zhezhen Jin
Chair: Yingwei Paul Peng

14:00 Fairness-Constrained Optimal Model Averaging with High-Dimensional Sparsity Learning

Zeyu Chen, ♦Wei Qian, Bintong Chen

University of Delaware University of Delaware University of Delaware

14:25 Leveraging Multiple Endpoints to Estimate and Identify Subgroup-Specific Treatment Effects

♦Tom Chen, Emma Smith, Nick Birk, Rui Wang

Harvard Medical School Harvard T. H. Chan School of Public Health Harvard T. H. Chan School of Public Health Harvard T. H. Chan School of Public Health & Harvard Medical School

14:50 A unified regression-based method for X-chromosome-inclusive Hardy–Weinberg equilibrium

♦Lin Zhang, Andrew Paterson, Lei Sun

Simon Fraser University, The Hospital for Sick Children University of Toronto

25CHI051: Modern Statistical Learning

Room: C304
Organizer: Lexin Li
Chair: Yin Xia

14:00 Statistical Learning via Partial Derivatives

♦Xiaowu Dai

University of California, Los Angeles

14:25 SAT: Data-light Uncertainty Set Merging via Synthetics, Aggregation, and Test Inversion

Shenghao Qin, Jianliang He, Bowen Gang, ♦Yin Xia

Fudan University, Yale University, Fudan University, Fudan University

14:50 Optimal PhiBE — A New Framework for Continuous-Time Reinforcement Learning

♦Yuhua Zhu

University of California, Los Angeles

15:15 On the Optimality of Inference on the Mean Outcome under Optimal Treatment Regime

Shuoxun Xu, ♦Xinzhou Guo

UC Berkeley, HKUST

25CHI056: New advances in statistical theory, method and application

Room: C307
Organizer: Le Zhou
Chair: Le Zhou

14:00 Analysis of Sparse Sufficient Dimension Reduction Models

Yeshan Withanage, Wei Lin, ♦Zhijian Li

JPMorgan Chase Bank Ohio University Beijing Normal - Hong Kong Baptist University

14:25 Two-fold Varying-coefficient Mediation Models and Their Applications

Jie Xing, Le Zhou, Tiejun Tong, ♦WenWu Wang

Qufu Normal University Hong Kong Baptist University Hong Kong Baptist University Qufu Normal University

14:50 When Tukey meets Chauvenet: a new boxplot criterion for outlier detection

♦Hongmei Lin, Riquan Zhang, Tiejun Tong

Shanghai University of International Business and Economics Shanghai University of International Business and Economics Hong Kong Baptist University

15:15 Structural Testing of High-dimensional Correlation Matrices

♦Tingting Zou, Guangren Yang, Ruitao Lin, Guo-Liang Tian, Shurong Zheng

Jilin University Jinan University The University of Texas MD Anderson Cancer Center Southern University of Science and Technology Northeast Normal University

25CHI057: New frontiers in large-scale data analysis with applications to heterogeneous data

Room: C305
Organizer: Yichuan Zhao
Chair: Yichuan zhao

14:00 Kernel Regression Utilizing External Information as Constraints

♦Chi-Shian Dai, Shao Jun

National Cheng Kung Univeristy, University of Wisconsin-Madison

14:25 An alternative measure for quantifying the heterogeneity in meta-analysis

Ke Yang, Enxuan Lin, Wangli Xu, Liping Zhu, ♦Tiejun Tong

Beijing University of Technology Innovent Biologics, Inc. Renmin University of China Renmin University of China Hong Kong Baptist University

14:50 Statistical inference for large-scale multi-source heterogeneous data

Jiuzhou Miao, ♦Li Cai, Suojin Wang

Zhejiang Gongshang university Zhejiang Gongshang University Texas A&M university

15:15 Power Enhancement Subsampling Adaptive Ensemble Test

⧫*Xuehu Zhu*

Xi'an Jiaotong University

25CHI062: On Statistical Stability and Ensemble Learning

Room: C306
Organizer: Wei Zhong
Chair: Wei Zhong

14:00 Integrating Inference Results via Synthetic Statistics.

Shenghao Qin, Jianliang He, ⧫*Bowen Gang, Yin Xia*

Fudan University, Yale University, Fudan University, Fudan University

14:25 A Gaussian Stability framework for Post-Selection Inference

Jiajun Sun, Chendi Wang, ⧫*Wei Zhong*

Xiamen University Xiamen University Xiamen University

14:50 A simple and powerful method for large-scale composite null hypothesis testing with applications in mediation analysis

⧫*Yaowu Liu*

Southwestern University of Finance and Economics

f. June 29th PM (16:00-17:40)

25CHI066: Recent Advances in Correcting Measurement Error in Epidemiological Research

Room: C404
Organizer: Xin Zhou
Chair: Xin Zhou

16:00 Causal inference for the effects of mismeasured covariates through double/debiased machine learning

Gang Xu, Xin Zhou, Molin Wang, Boya Zhang, Donna Spiegelman, ⧫*Zuoheng Wang*

Yale University, Yale University, Harvard University, Harvard University, Yale University, Yale University

16:25 Time-to-Event Analysis of Preterm Birth Accounting for Gestational Age Uncertainties

⧫*Yuzi Zhang, Joshua Warren, Hua Hao, Howard Chang*

Division of Biostatistics, The Ohio State University Department of Biostatistics, Yale University Department of Environmental Health, The Rollins School of Public Health of Emory University Department of Biostatistics and Bioinformatics, Emory University

16:50 Survival analysis adjusting for measurement error in a cumulative exposure variable: radon progeny to lung cancer mortality

⧫*Molin Wang, Yue Yang, Donna Spiegelman*

Harvard University, Harvard University, Yale University

17:15 Generalized Methods-of-Moments estimation and inference for the assessment of multiple imperfect measures of physical activity in validation studies

Zexiang Li, Donna Spiegelman, Molin Wang, Zuoheng Wang, ⧫*Xin Zhou*

Yale School of Public Health Yale School of Public Health Harvard T.H. Chan School of Public Health Yale School of Public Health Yale School of Public Health

25CHI067: Recent advances in high dimensional data and machine learning

Room:C203
Organizer: Xiaochao Xia
Chair: Xiaochao Xia

16:00 Privacy-Preserving Transfer Learning for Community Detection using Locally Distributed Multiple Networks

Xiao Guo, Xuming He, Xiangyu Chang, ⧫*Shujie Ma*

Northwest University of China, Washington University in St. Louis, Xi'an Jiaotong University, University of California-Riverside

16:25 Statistical inference for high-dimensional convoluted rank regression

Leheng Cai, ⧫*Xu Guo, Heng Lian, Liping Zhu*

Tsinghua University, Beijing Normal University, City University of Hong Kong, Renmin University of China

16:50 Clustering functional data with measurement errors: a simulation-based approach

Tingyu Zhu, ⧫*Lan Xue, Carmen Tekwe, Keith Diaz, Mark Benden, Roger Zoh*

Oregon State University, Oregon State University, Indiana University, Columbia University Medical Center, Texas A&M University, Indiana University

17:15 Fair classification with continuous sensitive attribute

⧫*Xianli Zeng, Edgar Dobriban*

Xiamen University

25CHI068: Recent Advances in Machine Learning Techniques for Point Process Models

Room:C204
Organizer: Biao Cai
Chair: Shizhe Chen

16:00 Score Matching for Point Processes

⧫*Feng Zhou*

Renmin University of China

16:25 Bayesian inference for independent cluster point processes

Marie-Colette van Lieshout, ♦Changqing Lu

National Research Institute for Mathematics and Computer Science in the Netherlands; University of Twente National Research Institute for Mathematics and Computer Science in the Netherlands; University of Twente

16:50 Residual TPP: A unified lightweight approach for event stream data analysis

Ruoxin Yuan, ♦Guanhua Fang

Fudan University, Fudan University

17:15 Tensor-based Estimation and Inference for High-Dimensional Multivariate Point Process

Xiwei Tang, ♦Gannggang Xu, Jingfei Zhang, Yongtao Guan

University of Texas at Dallas, University of Miami, Emory University, Chinese University of Hong Kong, Shenzhen

25CHI074: Recent Advances in Statistical and Machine Learning

Room:C205
Organizer: Chengchun Shi
Chair: Jin Zhu

16:00 Contextual Dynamic Pricing: Algorithms, Optimality, and Local Differential Privacy Constraints

Zifeng Zhao, ♦Feiyu Jiang, Yi Yu

University of Notre Dame, Fudan University, University of Warwick

16:25 Sequential knockoffs for variable selection in reinforcement learning

Tao Ma, ♦Jin Zhu, Hengrui Cai, Zhengling Qi, Yunxiao Chen, Chengchun Shi, Eric Laber

London School of Economics and Political Science, London School of Economics and Political Science, University of California, Irvine, George Washington University, London School of Economics and Political Science, London School of Economics and Political Science, Duke University

16:50 Joint robust estimation

♦Lihu Xu

University of Macau

17:15 Consistent Selection of the Number of Groups in Panel Models via Cross-Validation

Zhe Li, Changliang Zou, ♦Xuening Zhu

Fudan University

25CHI085: Recent developments of high dimensional model checking

Room:C405
Organizer: Falong Tan
Chair: Falong Tan

16:00 Testing mutually exclusive hypotheses for multi-response regressions

♦Jiaqi Huang, Wenbiao Zhao, Lixing Zhu

Beijing Normal University, China University of Mining & Technology, Beijing Beijing Normal University

16:25 Goodness-of-Fit Tests for High-Dimensional Regression Models via Projections

Wen Chen, Jie Liu, Heng Peng, ♦Falong Tan, Lixing Zhu

Hunan University, Hunan University, Hong Kong Baptist University, Hunan University, Beijing normal University

16:50 Model checking for parametric regressions in transfer learning

Chuhan Wang, Jiaqi Huang, ♦Xuerui Li

Beijing Normal University, Beijing Normal University, Beijing Normal University

17:15 A Chi-Square Specification Test with One-Class Support Vectors

Yuhao Li, ♦Xiaojun Song

Xi'an Jiaotong-Liverpool University, Peking University

25CHI086: Recent Progresses in Nonparametric and Semiparametric Statistics

Room:C207
Organizer: Ying Yan
Chair: Ying Yan

16:00 Enhanced Fused Sufficient Representation Learning for Neuroimaging Data

Yue Chen, Baiguo An, Linglong Kong, Xueqin Wang, ♦Wenliang Pan

Capital University of Economics and Business, Capital University of Economics and Business, University of Alberta, University of Science and Technology of China, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

16:25 Unsupervised optimal deep transfer learning for classification under general conditional shift

Junjun Lang, ♦Yukun Liu

East China Normal University, East China Normal University

16:50 Semi-parametric inference on inequality measures with non-ignorable non-response using callback data

♦Chunlin Wang

Xiamen University

17:15 Matching-based Policy Learning

♦Ying Yan

Sun Yat-sen University

25CHI090: Scalable Learning and Knowledge Transfer for Complex Biomedical Data

Room:C208
Organizer: Xinping Cui
Chair: Gang Li

16:00 Data-Driven Knowledge Transfer in Batch Q* Learning

Elynn Chen, Xi Chen, ♦Wenbo Jing

New York University, New York University, New York University

16:25 A transfer learning approach for interval-censored failure time data

♦ *(Tony) Jianguo Sun*

University of Missouri

16:50 Applications of Bayesian Power Prior with a Discount Function in Medical Device Trials

♦*Hong Zhao*

Abbott

17:15 Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Survival Outcome for Large Data

Shanpeng Li, Emily Ouyang, Jin Zhou, Xinping Cui, ♦Gang Li

City of Hope, UCR, University of California at Los Angeles, UCR, University of California at Los Angeles

25CHI095: Statistical Advances for Integrative Multi-Omics Data Analysis

Room:C209
Organizer: Jiebiao Wang
Chair: Jiebiao Wang

16:00 Local genetic correlation via knockoffs reduces confounding due to cross-trait assortative mating

♦*Shiyang Ma, Fan Wang, Iuliana Ionita-Laza*

Shanghai Jiao Tong University, Columbia University, Columbia University

16:25 scHiCPRSiM: single-cell Hi-C Practical and Rational SiMulator

♦*Huiling Liu, Rui Ma, Xingping Cui, Wenxiu Ma*

South China University of Technology

16:50 Spatial Resolved Gene Regulatory Networks Analysis

Ishita Debnath, ♦Zhana Duren

Indiana University, Indiana University

17:15 Clustering-free nuanced marker identification and attribution and its application in the taurine compensatory effect discovery in dilated cardiomyopathy

Jinpu Cai, Xiaorui Liu, Cheng Wang, Yibo Zhang, Luting Zhou, Ziqi Rong, Hongbin Shen, Qiuyu Lian, Liang Chen, ♦Hongyi Xin

Shanghai Jiao Tong University, Fuwai Hospital, Shanghai Jiao Tong University, Fuwai Hospital, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Cambridge University, Fuwai Hospital, Shanghai Jiao Tong University

25CHI102: Statistical learning based on high dimensional and complex data

Room:C210
Organizer: Ming-Yen Cheng
Chair: Jin-Ting Zhang

16:00 Statistical Approaches to MLP Approximation in Efficient Language Models

♦*Yifan Chen*

Hong Kong Baptist University

16:25 Local Information for Global Network Estimation in Latent Space Models

♦*Lijia Wang, Xiao Han, Yanhui Wu, Y. X. Rachel Wang*

City University of Hong Kong, University of Science and Technology of China, University of Hong Kong, University of Sydney

16:50 Statistical inference for functional data over multi-dimensional domain

Qirui Hu, ♦Lijian Yang

Tsinghua University, Tsinghua University

25CHI109: Statistical Network Analysis and Applications

Room:C301
Organizer: Binyan Jiang
Chair: Binyan Jiang

16:00 Modelling Homophily in Autoregressive Networks

♦*Xinyang Yu*

London School of Economics and Political Science

16:25 A network approach to compute hypervolume under receiver operating characteristic manifold for multi-class biomarkers

♦*Qunqiang Feng, Pan Liu, Pei-Fen Kuan, Fei Zou, Jianan Chen, Jialiang Li*

University of Science and Technology of China, National University of Singapore, Stony Brook University, University of North Carolina at Chapel Hill, National University of Singapore, National University of Singapore

16:50 Community detection in weighted networks via the profile-pseudo likelihood method

Yang Liu, Jiangzhou Wang, ♦Binghui Liu

Northeast Normal University, Shenzhen University, Northeast Normal University

17:15 Transfer Learning Under High-Dimensional Network

Convolutional Regression Model

Liyuan Wang, Jiachen Chen, Kathryn Lunetta, ♦Danyang Huang, Huimin Cheng, Debarghya Mukherjee

Renmin University of China

25CHI110: Statistical theory of neural networks

Room:C302
Organizer: Jun Fan
Chair: Jun Fan

16:00 Rates for least squares using over-parameterized neural networks

♦Yunfei Yang

Sun Yat-sen University

16:25 Solving PDEs on Spheres with Physics-Informed Convolutional Neural Networks

♦Lei Shi

Fudan University

16:50 Optimization and Generalization of Gradient Methods for Shallow Neural Networks

♦Yunwen Lei, Yiming Ying, Ding-Xuan Zhou

The University of Hong Kong, University of Sydney, University of Sydney

17:15 Nonparametric GARCH: A Deep Learning Approach

Ruizhi Deng, ♦Guohao Shen, Ngai Hang Chan

The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, The City University of Hong Kong

25CHI113: Theoretical Advances in Machine Learning and Dimension Reduction, and Functional Data Analysis

Room:C304
Organizer: Dongming Huang
Chair: Dongming Huang

16:00 On the structural dimension of sliced inverse regression

♦Dongming Huang, Songtao Tian, Qian Lin

National University, Tsinghua University, Tsinghua University

16:25 Diagonal Over-parameterization in Reproducing Kernel Hilbert Spaces as an Adaptive Feature Model: Generalization and Adaptivity

♦Yicheng Li, Qian Lin

Department of Statistics and Data Science, Tsinghua University, Department of Statistics and Data Science, Tsinghua University

16:50 Duality Between Context Data and Model Parameters in Transformers

Brian Chen, Hui Jin, Haonan Wang, Kenji Kawaguchi, ♦Tianyang Hu

National University of Singapore, Huawei Noah's Ark Lab, National University of Singapore, National University of Singapore, National University of Singapore

17:15 Modified Tests of Linear Hypotheses Under Heteroscedasticity for Multivariate Functional Data with Finite Sample Sizes

♦Tianming Zhu

National Institute of Education, Nanyang Technological University

25CHI037: Innovative Inference Methods for Complex Data: Bridging Theory and Practice

Room:C305
Organizer: Xingche Guo
Chair: Xingche Guo

16:00 A practical interval estimation method for spectral density function

♦Haihan Yu, Mark Kaiser, Daniel Nordman

University of Rhode Island, Iowa State University, Iowa State University

16:25 A Semiparametric Causal Estimator without Ignorability

♦Guoliang Ma, Cindy Yu, Zhonglei Wang

Xiamen University, Iowa State University, Xiamen University

16:50 Online Tensor Inference

♦Xin Wen, Will Wei Sun, Yichen Zhang

New York University, Purdue University, Purdue University

25CHI048: Modern Machine Learning: Tackling Real-World Data Challenges

Room:C306
Organizer: Xinping Cui
Chair: Xiaoqian Liu

16:00 Renewable l1-regularized linear support vector machine with high-dimensional streaming data

Na Zhang, Jinhan Xie, Xiaodong Yan, Bei Jiang, Ting Li, ♦Linglong Kong.

University of Alberta, Yunan University, Xi'an Jiaotong University, University of Alberta, Hong Kong Polytechnic University, University of Alberta

16:25 Inference of Comparing Generative AI Models

♦Zijun Gao, Yan Sun

University of Southern California, University of Pennsylvania

16:50 Transfer Reinforcement Learning: Value-Based Methods for Non-Stationary MDPs

♦Elynn Chen

New York University

17:15 Departments of Data Science and AI, and Applied Mathematics

♦Jian Huang

The Hong Kong Polytechnic University

25CHI049: Modern Multivariate Analysis for Tensor and Multiview Data

Room:C307
Organizer: Xin Zhang
Chair: Jing Zeng

16:00 D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High-dimensional Data

♦Hai Shu, Hongtu Zhu

New York University, The University of North Carolina at Chapel Hill

16:25 Low-rank decomposition, dimension-reduction subspaces and tensor change detection

Jiaqi Huang, ♦Ning Wang, Lixing Zhu

Beijing Normal University

16:50 Statistical Inference for Low-Rank Tensor Models

♦Ke Xu, Elynn Chen, Yuefeng Han

University of Notre Dame, New York University, University of Notre Dame

17:15 Sparse and integrative principal component analysis for multiview data

Lin Xiao, ♦Luo Xiao

North Carolina State University, North Carolina State University

g. June 30th AM (8:30-10:10)

25CHI013: Advances in Statistical Learning and Network Analysis for Complex Data

Room: C203
Organizer: Peng Liu
Chair: Xiaofei Zhang

8:30 Adaptive Block-Based Change-Point Detection for Sparse Spatially Clustered Data with Applications in Remote Sensing Imaging

Alan Moore, Lynna Chu, ♦Zhengyuan Zhu

Iowa State University, Iowa State University, Iowa State University

8:55 Information-incorporated Network Construction with FDR Control

Hao Wang, Yumou Qiu, ♦Peng Liu

Iowa State University, Peking University, Iowa State University

9:20 Bias-corrected Byzantine-robust Estimator via Cornish-Fisher Expansion for Distributed Learning

Zhixiang Zhou, Yibo Yuan, Xiaojun Mao, ♦Zhonglei Wang

Xiamen University, Xiamen University, Shanghai Jiao Tong University, Xiamen University

9:45 Decentralized federated learning with fused lasso under distribution shift

Weidong Liu, Xiaojun Mao, ♦Xiaofei Zhang, Xin Zhang

Shanghai Jiao Tong University, Shanghai Jiao Tong University, Zhongnan University of Economics and Law, Meta

25CHI020: Advancing Risk Management with Statistical Learning

Room: C204
Organizer:Fan Yang
Chair: Fan Yang

8:30 Exploratory Investment-Consumption with Non-Exponential Discounting

Yuling Chen, ♦Bin Li, David Saunders

University of Waterloo, University of Waterloo, University of Waterloo

8:55 Optimal Pooling of Catastrophe Risks

Minh Chau Nguyen, Tony Wirjanto, ♦Fan Yang

University of Waterloo, University of Waterloo, University of Waterloo

9:20 Defense Against Syntactic Textual Backdoor Attacks with Token Substitution

Xianwen He, Xinglin Li, Minhao Cheng, ♦Yao Li

University of North Carolina at Chapel Hill, University of North Carolina at Chapel Hill, Pennsylvania State University, University of North Carolina at Chapel Hilll

9:45 The Joint Law of the Terminal Value, Running Maximum and Running Minimum of a Scalar Diffusion Process with Time-Inhomogeneous Drift

Philip Ernst, ♦Jixin Wang

Imperial College London, Imperial College London

25CHI032: Innovations in High Dimensional Complex Data Analysis: From Functional Data Analysis to Measurement Error Modeling

Room: C207
Organizer:Juan Xiong
Chair: Juan Xiong

8:30 Addressing Misclassification in Outcome and Covariate via A Likelihood-Based Approach

Zhegn Yu, ♦Hua Shen

University of Calgary, University of Calgary

8:55 High dimensional Recurrent Event Analysis with

Error-Contaminated Covariates

♦*Kaida Cai*

Southeast University

9:20 Generalized SIMEX Method: Polynomial Approximation for Extrapolation

Li-Pang Chen, ♦*Qihuang Zhang*

National Chengchi University, McGill University

9:45 Meta-analyzing multiple functional data with functional fixed-effects model

Jiahao Tang, ♦*Zongliang Hu, Hanbing Zhu, Yan Zhou, Shurong Zheng*

Northeast Normal University, Shenzhen University, Anhui University, Shenzhen University, Northeast Normal University

25CHI035: Innovative Approaches in Electronic Health Record (EHR) Data Analysis

Room: C208
Organizer:Molei Liu
Chair: Molei Liu

8:30 Age-Specific Outcome-guided Representation Learning for Patient Clustering with EHR Data

♦*Linshanshan Wang, Mengyan Li, Molei Liu, Zongqi Xia, Tianxi Cai*

Harvard University, Bentley University, Peking University, University of Pittsburgh, Harvard University

8:55 An Evaluation Framework for Ambient Digital Scribing Tools in Clinical Applications

Haoyuan Wang, Rui Yang, Mahmoud Alwakeel, Ankit Kayastha, Anand Chowdhury, Joshua M. Biro, Anthony D. Sorrentino, Michael J. Pencina, Kathryn I. Pollak, ♦*Chuan Hong*

Duke University, Duke-NUS Medical School, Duke University, Duke University, Duke University, Medstar Health National Center for Human Factors in Healthcare, Duke University, Duke University, Duke University, Duke University

9:20 Improving Robustness of the Model-X Inference with Application to EHR Studies

♦*Molei Liu*

Peking University

25CHI055: New advances in design and analysis of longitudinal studies

Room: C209
Organizer:Zhigang Li
Chair: Zhigang Li

8:30 Integrating multiple imperfect measures for alcohol use in longitudinal research studies

♦*Robert Cook, Samuel Wu, Donald Porchia, Yan Wang, Zhigang Li*

University of Florida, University of Florida, University of Florida, University of Florida, University of Florida

8:55 Integrating wearable sensors into longitudinal cohort studies: Opportunities and challenges

♦*Yan Wang*

University of Florida

9:20 Joint modeling in presence of informative censoring on the retrospective time scale

Quran Wu, Michael Daniels, Areej El-Jawahri, Marie Bakitas, ♦*Zhigang Li*

University of Florida, University of Florida, Harvard University, UAB, University of Florida

9:45 Correlation Coefficients for a Study with Repeated Measures

♦*Guogen Shan*

University of Florida

25CHI071: Recent Advances in Nonparametric Estimation and Inference

Room: C210
Organizer:Qing Wang
Chair: Qing Wang

8:30 Jackknife empirical likelihood for the correlation coefficient with multiplicative distortion measurement errors

Brian Pidgeon, Pangpang Liu, ♦*Yichuan Zhao*

Georgia State University, Purdue University, Georgia State University

8:55 Estimation of Multiple Large Precision Matrices and Its Application to High-Dimensional Quadratic Discriminant Analysis

Yilei Wu, Liyuan Zheng, ♦*Yingli Qin, Mu Zhu, Weiming Li*

University of Waterloo

9:20 A Unified Framework of Classification-based Equality Test of Distributions

Zhen Zhang, ♦*Xin Liu*

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

9:45 Minimax optimal two-stage algorithm for moment estimation under covariate shift

Zhen Zhang, Xin Liu, ♦*Shaoli Wang, Jiaye Teng*

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

25CHI072: Recent advances in privacy-protected data collection and analysis

Room: C302

Organizer:Samuel Wu
Chair: Zhigang Li

8:30 Learning from vertically distributed data across multiple sites: An efficient privacy-preserving algorithm for Cox proportional hazards model with variable selection

Guanhong Miao, Lei Yu, Jingyun Yang, David Bennett, Jinying Zhao, ♦Samuel Wu

University of South Florida, Rush University Medical Center, Rush University Medical Center, Rush University Medical Center, University of South Florida, University of South Florida

8:55 Distributed Proportional Likelihood Ratio Model With Application to Data Integration Across Clinical Sites

♦Jiasheng Shi, Chongliang Luo

The Chinese University of Hong Kong, Shenzhen, Washington University in St. Louis

9:20 Logistic Regression Model for Differentially-Private Matrix Masking Data

♦Linh Nghiem, Aidong Adam Ding, Samuel Wu

University of Sydney, Northeastern University, University of South Florida

25CHI103: Statistical Learning for Complex Data Structures

Room: C304
Organizer:Ruiyang Wu
Chair: Yuchen Zhou

8:30 Estimation and Inference for CP Tensor Factor Model

Bin Chen, ♦Yuefeng Han, Qiyang Yu

University of Rochester, University of Notre Dame, University of Rochester

8:55 Heteroskedastic Tensor Clustering

♦Yuchen Zhou, Yuxin Chen

University of Illinois Urbana-Champaign, University of Pennsylvania

9:20 Interpretable Classification of Categorical Time Series Using the Spectral Envelope and Optimal Scalings

♦Zeda Li, Scott Bruce, Tian Cai

Baruch College, CUNY, Texas A&M University, The City University of New York

9:45 Dynamic Supervised Principal Component Analysis for Classification

Wenbo Ouyang, ♦Ruiyang Wu, Ning Hao, Hao Zhang

University of Arizona, Baruch College, CUNY, University of Arizona, University of Arizona

25CHI104: Statistical learning with complex data

Room: C305
Organizer:Peter Song
Chair: Xinyuan Song

8:30 Lessons learned from LLM benchmarking & evaluation

♦Youna Hu

Amazon

8:55 Decentralized TD Learning with Spatio-temporal Information Dependence

♦Shaogao Lv

Nanjing Audit University

9:20 Quantile tensor factor regression with interaction effects and its application to multimodal data analysis

Pengfei Pi, ♦Shan Luo

Shanghai Jiao Tong University, Shanghai Jiao Tong University

9:45 Robust group detection and membership prediction

Boyan Shen, ♦Xuerong Chen, Yong Zhou

Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, East China Normal University

h. June 30th AM (10:30-12:10)

25CHI039: Innovative Statistical and Machine Learning Methods for Complex Health Data

Room: C203
Organizer:Jingjing Zou
Chair: Todd Ogden

10:30 Transformer-Based Self-Supervised Learning for Multimodal Wearable Data

♦Jingjing Zou

UC San Diego

10:55 Neyman Smooth-Type Goodness of Fit in Complex Surveys

Lang Zhou, ♦Yan Lu, Guoyi Zhang, Ronald Christensen

AbbVie Inc., University of New Mexico, University of New Mexico, University of New Mexico

11:20 Bayesian monotone regression with large number of covariates and complex structure

♦Ken Cheung, Keith Diaz

Columbia University, Columbia University

11:45 Functional Fixed and Random Effects Inference with Applications to Accelerometry Data

♦Erjia Cui

University of Minnesota

25CHI053: Modern Statistical Modeling in Medical Research

with Real World Data

Room: C204
Organizer:Zheyu Wang
Chair: Jing Huang

10:30 Time-Since-Infection Model for Hospitalization and Incidence Data

Jiasheng Shi, Yizhao Zhou, ♦Jing Huang

The Chinese University of Hong Kong, Shenzhen, AstraZeneca, University of Pennsylvania

10:55 Causal inference for all: Estimands of practical interest in intercurrent event settings

♦Ruixuan Zhao, Linbo Wang, Mats J. Stensrud

University of Toronto Scarborough, University of Toronto Scarborough, Ecole Polytechnique Fédérale de Lausanne

11:20 A functional spatial partitioning approach to lesion segmentation using MRIs

♦Lin Zhang

University of Minnesota

25CHI054: New Advancements in Statistical Learning

Room: C207
Organizer:Zheyu Wang
Chair: Yaofang Hu

10:30 Neural network on interval-censored data with application to the prediction of Alzheimer's disease

♦Tao Sun, Ying Ding

Renmin University of China, University of Pittsburgh

10:55 Variational Bayesian Semi-supervised Keyword Extraction

♦Yaofang Hu, Yichen Cheng, Yusen Xia, Xinlei Wang

University of Alabama, Georgia State University, Georgia State University, University of Texas at Arlington

11:20 Diffusion model for large spatial temporal data

♦Xin Tong

National University of Singapore

11:45 Consistent Order Determination of Markov Decision Process

♦Chuyun Ye, Lixing Zhu, Ruoqing Zhu

Beijing Normal University, Beijing Normal University at Zhuhai, University of Illinois at Urbana-Champaign

25CHI058: New statistical methods in nonlinear regression analyses

Room: C208
Organizer:Peter Song
Chair: Xuerong Chen

10:30 Regression Analysis of Semiparametric Cox-Aalen Transformation Models with Partly Interval-Censored Data

Xi Ning, ♦Yanqing Sun, Yinghao Pan, Peter Gilbert

Colby College, University of North Carolina at Charlotte, University of North Carolina at Charlotte, University of Washington and Fred Hutchinson Cancer Center

10:55 Collaborative quantile treatment effect estimation

Ye Fan, Ying Wei, Sung Nok Chiu, Tiejun Tong, ♦Nan Lin

Capital University of Economics and Business, Columbia University, Hong Kong Baptist University, Hong Kong Baptist University, Washington University in St. Louis

11:20 clusterMLD: An Efficient Clustering Method for Multivariate Longitudinal Data

Junyi Zhou, ♦Ying Zhang, Wenzhou Zhou

Amgen Inc., University of Nebraska Medical Center, Indiana University

11:45 Joint mixed membership modeling of multivariate longitudinal and survival data

Yuyang He, ♦Xinyuan Song, Kai Kang

The Chinese University of Hong Kong, The Chinese University of Hong Kong, Sun Yat-sen University

25CHI060: Novel statistical methods for complex data analysis

Room: C209
Organizer:Yichuan Zhao
Chair: Yichuan zhao

10:30 A Joint Model for Multiple Longitudinal Data with Different Missing Data Patterns and with Applications to HIV Prevention Trials

♦Jing Wu, Ming-Hui Chen, Jeffrey Fisher

University of Rhode Island, University of Connecticut, University of Connecticut

10:55 Model identification and selection for varying-coefficient EV models with missing responses

♦Mingtao Zhao, Houwu Wu, Fanqun Li

Anhui University of Finance and Economics, Anhui University of Finance and Economics, Anhui University of Finance and Economics

11:20 A Self-Normalized Two-Sample Test for Nonaligned Time Series with Heavy Tails and Long Memory

Weiliang Wang, Yu Shao, ♦Ting Zhang

Boston University, Boston University, University of Georgia

11:45 Neural frailty machines for survival analysis

Jiawei Qiao, Ruofan Wu, Guanhua Fang, ♦Wen Yu, *Zhiliang Ying*

Fudan University, Ant Group, Fudan University, Fudan University, Columbia University

25CHI077: Recent Advences in Statistical Learning for Biological and Biomedical Data

Room: C210
Organizer:Ruiyang Wu
Chair: Ruiyang Wu

10:30 Contrastive Learning on Multimodal Analysis of Electronic Health Records

♦Doudou Zhou, Tianxi Cai, Feiqing Huang, Ryumei Nakada, Linjun Zhang

National University of Singapore, Harvard T.H. Chan School of Public Health, Harvard T.H. Chan School of Public Health, Rutgers University, Rutgers University

10:55 Convex Covariate-adjusted Gaussian Graphical Regression

Ruobin Liu, ♦Guo Yu

University of California, Santa Barbara

University of California, Santa Barbara

11:20 Instrumental variable analysis with multivariate point process treatments.

Yu Liu, Zhichao Jiang, ♦Shizhe Chen

University of North Carolina, Chapel Hill, Sun Yat-Sen University, University of California, Davis

25CHI107: Statistical methods for the analysis of complex data

Room: C302
Organizer:Zhezhen Jin
Chair: Antai Wang

10:30 Targeted Inference for High-Dimensional Quantile Regression Models

Yakun Liang, Xuejun Jiang, ♦Jiancheng Jiang

Southern University of Science and Technology, Southern University of Science and Technology, University of North Carolina at Charlotte

10:55 Semiparametric Accelerated Failure Time Cure Model for Clustered Survival Data

Yi Niu, Duze Fan, Jie Ding, ♦Yingwei Peng

Dalian University of Technology, Dalian University of Technology, Dalian University of Technology, Queen's University

11:20 Robust causal effect estimation in high dimensional survival analysis via nonparametric learning

♦Shanshan Ding, Zhezhen Jin

University of Delaware, Columbia University

11:45 On Data-Enriched Logistic Regression

Cheng Zheng, Sayan Dasgupta, Yuxiang Xie, Asad Haris, ♦Yingqing Chen

University of Nebraska Medical Center, Fred Hutchinson Cancer Center, University of Washington, University of Washington, Stanford University

Abstracts of Invited Sessions

25CHI004: Advanced Statistical Methods and Applications in Medical and Image Data Analysis

High dimensional proteomics data added value in heart failure patient phenomapping

♦*Yinggan Zheng, Cindy Westerhout, Paul Armstrong*

University of Alberta, University of Alberta, University of Alberta

Background: Phenomapping applies statistical learning techniques to explore data patters thereby identifying distinct patient subgroups. Multidimensional, clinically rich patient information collected in contemporary clinical trials allows for phenomapping on a heterogeneous entity such as patients with heart failure. However, the incremental value of high dimensional proteomics data recently available in clinical trials is unclear.

Methods: Using five domains of patient-level data (clinical characteristics, electrocardiographic, echocardiographic, quantitative biomarker and targeted proteomics), we performed an agglomerative hierarchal clustering analysis to define phenogroups in a subset of a contemporary clinical trial of heart failure patients. We used the average silhouette width to evaluate cluster stability as well as determine the optimized number of phenogroups. The optimized number was also confirmed by the NbClust R package. We applied multinomial logistic regression models with stepwise selection to identify the most important variables in defining the phenogroups for the five domains respectively and also for all together. Reduction of Akaike information criterion (AIC) were calculated and compared in each model.

Results: A total of 105 variables on 564 participants were included in the clustering analysis. Three phenogroups were identified. Phenogroup 1 was young, well treated with guideline-directed medical therapy, and least likely to have an implantable cardioverter-defibrillator. Phenogroup 2 had the highest prevalence of atrial fibrillation and pathologic Q-waves on ECG. Phenogroup 3 was older, had more biventricular dysfunction, and advanced renal disease. When considering the conventional clinical characteristics domain alone, history of anemia, diabetes, and patient age, index event and NYHA class were identified as significant variables in defining the 3 phenogroups (Table; Model 1). Among Echocardiography domain model and ECG model, TAPSE and pathologic Q-waves were retained in Models 2 and 3, respectively. For the quantitative biomarkers model, GDF-15, cystatin C, and albumin were retained (Model 4). For the proteomics model, 9 proteins, including GDF-15, were remained as significant factors (Model 5), and this persisted in the final combined model (Model 6) eliminating other previously identified significant variables in Models 1–4. Relative to the null model, Model 6 had the largest reduction in AIC, reflecting superior fit (Model 6: 719.37; Model 4: 551.64; Model 3: 302.62; Model 2: 269.95; Model 1: 116.39).

Conclusion: Three distinct phenogroups were identified using clustering analysis without knowledge of outcomes. Proteomics data including GDF-15 showed most important contribution in defining these phenogroups. These findings may inform study entry criteria for future heart failure trials focused on the development of novel therapeutics.

Table. Identifying informative variables in the development of the heart failure phenogroups

Variables Wald

Chi-Square P value Reduction of AIC Compared with Null Model

Model 1: Clinical variables 116.39

History of Anemia 35.1568 <.0001

Diabetes 23.0476 <.0001

Age 22.8863 <.0001

Index event* 16.0742 0.0029

NYHA 15.0265 0.0005

Model 2: Echocardiogram 269.95

RV-TAPSE 19.9034 <.0001

Model 3: ECG 302.62

Q waves 7.9752 0.018

Model 4: Quantitative biomarkers 551.64

Cystatin C 79.2115 <.0001

GDF-15, quantitative assay 55.1962 <.0001

Albumin 26.2939 <.0001

Model 5: Proteomics 719.37

GDF-15 33.4729 <.0001

TFF3 28.5575 <.0001

SHPS-1 21.8411 <.0001

JAM-A 21.6717 <.0001

PI3 20.0556 <.0001

MMP-3 18.4618 <.0001

NOTCH-3 17.6069 0.0002

PAI 17.3858 0.0002

OPG 13.6813 0.0011

Model 6: All significant variables considered: Same as Model 5

*Index event refers to qualifying HF event: HF hospitalization 3–6 months, HF hospitalization within 3 months, or intravenous diuretic for HF (without hospitalization) within 3 months.

AIC=Akaike information criterion; ECG=electrocardiograph; GDF-15=growth differentiation factor 15; JAM-A=junctional adhesion molecule A; MMP-3=matrix metalloproteinase-3; NOTCH-3=neurogenic locus notch homolog protein 3; NYHA=New York Heart Association; OPG=osteoprotegerin; PAI=plasminogen activator inhibitor; PI3=elafin; RV-TAPSE=right ventricular tricuspid annular plane systolic excursion; SHPS-1=tyrosine-protein phosphatase non-receptor type substrate 1; TFF3=trefoil factor 3.

Variational Bayesian Logistic Tensor Regression with Application to Image Recognition

Yunzhi Jin, ♦Yanqing Zhang, Niansheng Tang

Yunnan University, Yunnan University, Yunnan University

In recent years, image recognition method has been a research hotspot in various fields such as video surveillance, biometric identification, unmanned vehicles, human-computer interaction, and medical image recognition. Existing recognition methods often ignore structural information of image data or depend heavily on the sample size of image data. To address this issue, we develop a novel variational Bayesian method for image classification in a logistic tensor regression model with image tensor predictors by utilizing tensor decomposition to approximate tensor regression. To handle the sparsity of tensor coefficients, we introduce the multiway shrinkage priors for marginal factor vectors of tensor coefficients. In particular, we obtain a closed-form approximation to the variational posteriors for classification prediction based on the matricization of tensor decomposition. Simulation studies are conducted to investigate the performance of the proposed methodologies in terms of accuracy, precision and F1 score. Flower image data and chest X-ray image data are illustrated by the proposed methodologies.

Conditional inference for ultrahigh-dimensional additive hazards model

♦*Meiling Hao*

University of International Business and Economics

In the realm of high-throughput genomic data, modeling with ultrahigh-dimensional covariates and censored survival outcomes is of great importance.We conduct conditional inference for the ultrahigh-dimensional additive hazards model, allowing both the covariates of interest and nuisance covariates to be ultrahigh-dimensional. The presence of right censorship with survival outcomes adds an extra layer of complexity to the original data structure, posing significant challenges for the ultrahigh-dimensional additive hazards model. To address this, we introduce an innovative test statistic based on the quadratic norm of the score function. Moreover, when there is a high correlation between the covariates of interest and nuisance covariates, we propose an orthogonalized score function-based test statistic to enhance statistical power. Additionally, we establish the limiting distribution of the test statistics under both the null and local alternative hypotheses, further enhancing the computational appeal of our approach. The proposed statistics are thoroughly evaluated through extensive simulation studies and applied to two real data examples.

Tensor-Based Individualized Treatment Rules for Neuroimaging Applications

♦*Yang Sui, Yuanying Chen, Ting Li, Yang Bai, Hongtu Zhu*

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, The University of North Carolina at Chapel Hill

Precision medicine aims to uncover the optimal personalized treatment rules, providing thoughtful decision support based on the characteristics of each patient. With the rapid advancement of medical imaging technology, incorporating patient-specific imaging information into individualized treatment rules has become increasingly critical. We introduce a novel, data-driven approach that utilizes both imaging data and additional covariates to guide the selection of optimal treatment strategies. Specifically, this study employs tensor and covariates within a regression framework, estimating optimal individualized treatment rules through Tucker decomposition. This method effectively reduces the number of parameters, leading to efficient estimation and computational feasibility. To handle high-dimensional tensors, we further employ sparse Tucker decomposition to further reduce the parameter space. We apply the proposed method to the ADNI dataset to identify both treatment-free and treatment-related effects of the covariates, including demographic and genetic features, and the baseline hippocampal surface, on future cognitive scores. Analysis results suggest that incorporating hippocampal imaging data significantly enhances predictive performance and supports better treatment assignment compared to using covariate information alone. Moreover, there is clear heterogeneity between the treatment-free coefficients and treatment-related coefficients corresponding to both the covariate and imaging data. Beyond the ordinary treatment-free effects, AD-related medications exert additional significant effects on cognitive MMSE scores through the presubiculum, with a smaller portion in the CA1 region of the right hippocampus.

25CHI005: Advanced Statistical Methods for analyzing health-related data

High-dimensional covariate-augmented overdispersed poisson factor model

Wei Liu, ♦*Qingzhi Zhong*

Sichuan University, Jinan University

The current Poisson factor models often assume that the factors are unknown, which overlooks the explanatory potential of certain observable covariates. This study focuses on high dimensional settings, where the number of the count response variables and/or covariates can diverge as the sample size increases. A covariate-augmented overdispersed Poisson factor model is proposed to jointly perform a high-dimensional Poisson factor analysis and estimate a large coefficient matrix for overdispersed count data. A group of identifiability conditions are provided to theoretically guarantee computational identifiability. We incorporate the interdependence of both response variables and covariates by imposing a low-rank constraint on the large coefficient matrix. To address the computation challenges posed by nonlinearity, two high-dimensional latent matrices, and the low-rank constraint, we propose a novel variational estimation scheme that combines Laplace and Taylor approximations. We also develop a criterion based on a singular value ratio to determine the number of factors and the rank of the coefficient matrix. Comprehensive simulation studies demonstrate that the proposed method outperforms the state-of-the-art methods in estimation accuracy and computational efficiency. The practical merit of our method is demonstrated by an application to the CITE-seq dataset.

Heterogeneous change-point Effects in Longitudinal Data: An Application to Age-Related Cognitive Decline

Xiaoke Li, Boxian Wei, ♦*Guangyu Yang, Min Zhang*

Vanke School of Public Health, Tsinghua University, Vanke School of Public Health, Tsinghua University, Institute of Statistics and Big Data, Renmin University of China, Vanke School of Public Health, Tsinghua University

Understanding the onset of cognitive decline during aging and identifying heterogeneity across sub-

populations is crucial for developing targeted health management strategies to enhance cognitive health in older populations. However, existing literature lacks consensus on the precise timing of cognitive decline onset, which can be viewed as a change-point estimation problem. Although extensive change-point estimation methods have been developed in statistical literature, most studies focus on cross-sectional data, and existing methods for longitudinal data face numerical challenges. These limitations hinder the generalization of existing methods to accommodate

heterogeneous change-point effects in longitudinal studies. In this work, we introduce a semismooth change-point estimation method based on continuous piecewise linear models for analyzing longitudinal data. We rigorously establish the consistency and asymptotic normality of the proposed estimator. To enhance numerical performance, we develop a novel and computationally efficient Tri-ToSNR algorithm. Additionally, we propose a Wald test to assess heterogeneous change-point effects. We apply our method to the China Health and Retirement Longitudinal Study dataset and characterize a heterogeneous age-related cognitive decline between urban and rural residents.

Multivariable Mendelian Randomization Method accounting for complex correlated and uncorrelated pleiotropy

♦*Qing Cheng, Li Cao*

Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics, Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics

Mendelian randomization (MR) is a powerful tool that leverages genetic variants as instrumental variables to infer causal relationships between exposures and outcomes. While univariable MR is effective, it cannot perform mediation analysis or disentangle direct effects of correlated exposures. Multivariable MR (MVMR) overcomes these limitations by simultaneously incorporating multiple exposures, enabling the estimation of direct causal effects while mitigating biases arising from pleiotropy. In this study, we introduce an efficient and robust MVMR method, termed MVMR-CUE, to handle both correlated and uncorrelated horizontal pleiotropy. Through extensive simulations, we demonstrate that MVMR-CUE outperforms existing MVMR methods in terms of robustness and efficiency. We apply MVMR-CUE to investigate the causal effects of ten exposures on eight complex diseases using large-scale genome-wide association study summary datasets. Our findings reveal several noteworthy insights, including the direct causal effects of triglyceride (TG), low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C) on coronary artery disease. In contrast, the association of LDL-C with Type 2 diabetes was attenuated after accounting for HDL-C and TG, suggesting that the latter two lipid measures play a more prominent role in the pathogenesis of Type 2 diabetes.

Statistical Inference with Mixed-Effect Model for Covariate-Adaptive Randomized Experiments

♦*Yang Liu, Lucy Xia, Feifang Hu*

Renmin University of China, Hong Kong University of Science and Technology, The George Washington University

Covariate adjustment is crucial in covariate-adaptive randomized experiments to ensure valid inferences for treatment effects. While fixed-effect models are widely accepted, mixed-effect models have gained increasing popularity in settings with numerous strata, though their theoretical properties under covariate-adaptive randomization remain underexplored. In this work, we investigate the advantages of mixed-effect models under a linear outcome framework, demonstrating that the mixed-effect estimator achieves a smaller variance for treatment effects compared to fixed-effect approaches when the sample size is small relative to the number of strata. This variance reduction directly depends on the marginal imbalance obtained from randomization, implying that designs optimizing finer balance yield more precise estimates and enhanced statistical power. For a sufficiently large sample size relative to the number of strata, we show that the treatment effect estimates obtained from fixed-effect and mixed-effect models are asymptotically equivalent. Our theoretical findings are validated through simulations and a clinical trial case study, highlighting the advantage of mixed effect model and covariate-adaptive randomization on the inference of treatment effect.

25CHI009: Advances in Complex Time Series and Spatial Modelling and Learning

Coefficient Shape Transfer Learning for Functional Linear Regression

♦*Shuhao Jiao, Ian Mckeague, Ngai-Hang Chan*

City University of Hong Kong, Columbia University, City University of Hong Kong

We develop a new transfer learning framework for functional linear models to address the challenge of data scarcity. The framework incorporates samples from the target model (target domain) alongside those from auxiliary models (source domains), transferring knowledge of coefficient shape from the source domains to the target domain. There are two major advantages to coefficient shape transfer: first, it is robust to covariate scaling, and second, coefficient shape homogeneity is more inclusive than coefficient homogeneity, enabling the framework to incorporate more useful information, thereby enhancing model estimation. We thoroughly investigate the convergence rate of the new estimator, improved by the transferred knowledge, and study the optimality of these rates. We find that the improvement in model estimation depends not only on the similarity of coefficient shapes between the target and source domains, but also on the coefficient magnitude and the spectral decay rate of the covariance operators of the functional covariates in these domains. Additionally, to address the case where only a subset of the auxiliary models is informative for the target model, we develop an identification procedure to select the informative auxiliary models.

Nonparametric Estimation of Weakly Dependent Time Series via Neural Networks

Zudi Lu, Shubin Wu, ♦*Gan Yuan, Chao Zheng*

City University of Hong Kong, University of Southampton, City University of Hong Kong, University of Southampton

We consider the non-parametric regression via neural networks. Despite the great empirical success deep learning has achieved in the past years, most of its theoretical development relies on i.i.d. assumption. In this work, we relax the independent assumption by allowing weakly dependent observations. In particular, we show that a fully connected neural network can achieve the minimax generalization error bound for smooth regression functions up to a polylogarithmic factor, when the observations are assumed to be alpha-mixing. Furthermore, we investigate how the network architectures (e.g., depth and width) affect the prediction accuracy. A comprehensive simulation study is conducted to evaluate the empirical performance of neural network prediction with various levels of weak dependency.

A Root-n-Consistent Semiparametric Superquantile Autoregression for Dynamic Time Series with a Possibly Incorrect Model Specification

Jiangtao Wang, ♦Zudi Lu, Xiyu Zhou, Wu Jin

School of Economics and Business Administration, Central China Normal University, China, Department of Biostatistics, City University of Hong Kong, Hong Kong SAR, China, School of Economics and Business Administration, Central China Normal University, China, School of Economics and Business Administration, Central China Normal University, China

Extending autoregressive quantile, we propose two schemes of semiparametric autoregressive superquantile (SQ), also called expected shortfall (ES) in risk management, under possibly incorrect model specifications. These schemes are inspired by a nonlinear autoregressive location-scale (LOSC) framework, extending the quantile ARMA and GARCH models. The SQ and quantile autoregressions jointly approximated optimally with their estimators are suggested via an elicitable universal loss. They exhibit root-n asymptotic normalities, but with varied variances depending on whether the LOSC-based superquantile and quantile hold true or are optimally approximated otherwise. For a confidence interval for prediction, we propose a random weight resampling method that overcomes the challenging bootstrap for dependent time series data with a possibly incorrect model. This resampling is appealing as it is both model and data-free with the resampling-induced estimators automatically adapting to the varied asymptotic variances. The methodology is supported by finite-sample Monte Carlo simulations and an application to expected shortfall forecasting for real data.

Covariance parameter estimation for spatial models

♦Saifei Sun

City University of Hong Kong

The purpose is to consider the covariance parameter estimation for Gaussian random fields that are observed with measurement error and irregularly spaced design sites on a fixed and bounded domain. The Gaussian random fields are assumed to have smooth mean functions and isotropic covariance functions belonging to powered exponential, Matrn, or generalized Wendland class. Under fixed-domain asymptotics, consistent estimators are proposed for three microergodic parameters, namely the nugget, the smoothness parameter, and a parameter related to the coefficient of the principal irregular term of the covariance function. Upper bounds for the convergence rate of these estimators are also established.

25CHI012: Advances in Statistical Learning and Algorithm

Optimal Model Averaging for Imbalanced Classification

♦Ze Chen, Jun Liao, Wangli Xu, Yuhong Yang

Shandong University, Renmin University of China, Renmin University of China, Tsinghua University

Imbalanced data with a high-dimensional input has been widely encountered in many areas of applications. In this situation, it usually becomes

essential to reduce redundant variables via model selection to improve the classification performance. However, with a large number of variables, model selection uncertainty is typically very high. To deal with this problem, we present a feasible model averaging procedure based on a cost-sensitive support vector machine (CSSVM) coupled with a cost-sensitive data-driven weight choice criterion for imbalanced classification. Theoretical justifications are provided in two distinct scenarios. When the data exhibits a weak imbalance, we derive a relatively fast uniform convergence rate of the CSSVM solution. In contrast, when the data possesses a strong imbalance, the convergence rate becomes much slower. In both scenarios, an asymptotic optimality of the proposed model averaging approach in the sense of minimizing the out-of-sample hinge loss is established. Moreover, to reduce the computational burden imposed by a large number of candidate models for model averaging, we develop the CSSVM with an L1-norm penalty to prepare candidate models. Numerical analysis shows the superiority of the proposed model averaging procedure over existing imbalanced classification methods.

A paradox in Metropolis-Hastings practice

♦Jiandong Shi, Sheng Lian, Xiaodan Fan

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

Markov Chain Monte Carlo (MCMC) includes a class of powerful algorithms to generate sequential samples from a target distribution, where the Metropolis-Hastings (MH) algorithm is among the most popular ones. The MH algorithm, by its acceptance rule for the proposal point, shows the preference of the point of larger target distribution values. The intuition, though correct for low-dimensional settings, is broken down by a paradox when the dimension gets high. It has been observed from two common examples for applying MH algorithm that through the MCMC chain, the target distribution takes a large value at the initial point, decreases to a low density region, and gets stable.

This indicates that the MH algorithm keeps accepting proposal points of lower densities and the mode with the highest density may even not appear after the chain converges, which is quite counter-intuitive. We explain in detail how and why they happen via theoretical analysis and numerical studies. Furthermore, this paradox also raises a question whether the mode can be reasonably traced by the MH algorithm. Three common models are further analyzed to discuss about the reliability of the posterior mode in Bayesian analysis.

The Binary and Ternary Quantization Can Improve Feature Discrimination

♦Weiyu Li, Weizhi Lu, Mingrui Chen

Shandong University, Shandong University, Shandong University

In machine learning, quantization is widely used to simplify data representation and facilitate algorithm deployment on hardware. Considering the fundamental role of classification in machine learning, it is imperative to investigate the impact of quantization on classification. Current research primarily revolves around quantization errors, under the assumption that higher quantization errors generally lead to lower classification performance. However, this assumption lacks a solid theoretical foundation, and often contradicts empirical findings. For instance, some extremely low bit-width quantization methods, such as {0,1}-binary quantization and {0,-1,1}-ternary quantization, can achieve comparable or even superior classification accuracy compared to the original non-quantized data, despite exhibiting high quantization errors. To evaluate the classification performance more accurately, we propose to directly investigate the feature discrimination of quantized data, rather than analyze its quantization error. It is found that binary and ternary quantization can surprisingly improve, rather than degrade, the feature discrimination of original data. This remarkable performance is validated through classification experiments on diverse data types, including images, speech and text.

Quantile-Matched DC in Massive Data Regression

Yan Chen, ♦Lu Lin

Shandong University, Shandong University

The issues of bias-correction and robustness are crucial in the strategy of divide-and conquer (DC) for asymmetric nonparametric regression with massive data sets. Instead of single-quantile regression as used in the common DC methods, a multi-quantile DC is proposed in this paper to establish a weighted composite quantile regression (WCQR). By fully utilizing the model structure and elaborating the weight matched to the multi-quantile level, the WCQR can eciently correct the non-negligible bias of the nonparametric quantile regression with asymmetric error distribution. Moreover, in the procedure of correcting the non-negligible bias, the selected weights are robust via converting the bias into a parametric form, without use of any nonparametric pilot estimations. Then, the whole estimation process is robust and adaptive to the outliers, the fat-tailed distributions and the heterogeneity of data in dierent data batches. The theoretical properties of the new methods are systematically investigated and the behaviors are further illustrated by comprehensive simulation studies and real data analyses. Compared with the competitors, the new methods have the favorable features of estimation accuracy, robustness, applicability and communication efficiency.

25CHI016: Advances in Statistical Methods for Complex Data Integration and Causal Inference

Deep Clustering Evaluation: How to Validate Internal Clustering Validation Measures

Zeya Wang, ♦Chenglong Ye

University of Kentucky, University of Kentucky

Deep clustering partitions complex high-dimensional data using deep neural networks for clustering. It involves projecting data into lower-dimensional embeddings before partitioning, which embarks unique evaluation challenges. Traditional clustering validation measures, designed for low-dimensional spaces, are problematic for deep clustering for two reasons: 1) the curse of dimensionality when applied to the high-dimensional input data, and 2) unreliable comparison of clustering results when applied to embedded data from different embedding spaces, owing to variations in training procedures and model parameter settings. This paper addresses these unresolved and often overlooked challenges in evaluating clustering within deep learning. We propose a systematic evaluation framework for internal clustering validation measures that: (1) theoretically establishes why traditional measures are ineffective when applied to input data or across disparate embedding spaces paired with partitioning outcomes; (2) identifies embedding spaces that endorse reliable evaluations by detecting groups with high agreement in ranking partitioning outcomes; and (3) develops a stable and robust scoring scheme by weighting index values computed across these identified embedding spaces. Experiments show that this new framework aligns better with external measures, effectively reducing the misguidance from the improper use of internal validation measures in deep clustering evaluation.

Robust High-dimensional Inference for Causal Effects Under Unmeasured Confounding and Invalid Instruments with an Application to Multivariable Mendelian Randomization Analysis

♦Yunan Wu, Lan Wang, Baolin Wu, Yixuan Ye, Hongyu Zhao

Tsinghua University, University of Miami, UC Irvine, Yale University, Yale University

We consider a novel high-dimensional robust estimation and inference procedure for the causal effects in the presence of unmeasured confounding and invalid instruments based on observational data. Compared with the existing literature on causal inference using instrumental variables, our approach has several distinctive features. We do not assume the prior knowledge of a set of relevant instruments. The uncertainty of the availability of such a set is built into the inference procedure. In fact, our framework allows for the simultaneous violation of any of the three commonly imposed instrument validity conditions. We also allow the measured confounders to be endogenous. Our condition for the identification of causal effects, estimation and inference procedures do not require the specification of an exposure model. In particular, our method allows for a nonlinear relationship among the exposure, the instruments and other variables. The proposed inference procedure allows for high-dimensional instruments and/or high dimensional measured confounders. Our new procedure exploits the sparsity of the observed data model to identify the causal effects with potentially invalid instruments or many weak instruments. The validity of the confidence intervals is established under relatively weak conditions without requiring prior knowledge of a subset of valid instruments. We consider as a prime example Mendelian Randomization (MR) analysis with genetic instruments, multivariate exposures, and both measured and unmeasured confounders, based on individual-level data. We demonstrate in Monte Carlo studies that our new method has satisfactory performance and is robust to invalid instruments. We also illustrate the usefulness of our method through its application to the UK Biobank.

High-Resolution Feature Identification in High-Dimensional Clustering

♦Lyuou Zhang

Shanghai University of Finance and Economics

Interpretable clustering, which involves simultaneous identifying heterogeneous subpopulations and the informative features that define the subpopulations, is a critical yet challenging task across various fields, including omics studies, clinical research, and policy evaluation. Existing methods typically either focus narrowly on global feature heterogeneity or treat feature identification and clustering as separate tasks, failing to account for their interaction. To address these limitations, we propose a novel unsupervised learning approach, PAirwise REciprocal fuSE

(PARSE), which concurrently pinpoints cluster-specific informative features and conducts high-dimensional clustering effectively.

Combining variable screening methods for model averaging in High-Dimensional Data Analysis

♦*Zhihao Zhao, Yuhong Yang, Li Wen*

Capital University of Economics and Business, Tsinghua University, Renmin University of China

Both model averaging (MA) and variable screening have been sub- jects of extensive research. The literature has introduced various methods for variable screening, typically relying on a single approach, such as marginal correlation. However, the performances of such methods may be very poor out- side of their specifically applicable scenarios, and it is often difficult to know which scenarios are proper in real data analysis. In this study, we propose a method called Recommendation and Trimmed Mean Scoring (RTMS) for vari- able screening, which enhances reliability and stability of variable screening by integrating rankings from multiple screening methods. We introduce a practi- cal concept of approximate consistency in ranking of variables, establishing a theoretical property of the RTMS method in both hard sparse and gradually decaying coefficient scenarios. Simulation and empirical results demonstrate that the RTMS method outperforms individual ranking methods in terms of variable screening and predictive performance of MA as well.

25CHI022: Complex Structured Data Analysis

High-dimensional large-scale mixed-type data imputation under missing at random

♦*Wei Liu, Guizhen Li, Ling Zhou, Lan Luo*

School of Mathematics, Sichuan University, School of Economics and Finance, Guizhou University of Commerce, Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Department of Biostatistics and Epidemiology, Rutgers University

Missingness in mixed-type variables is commonly encountered in a variety of areas. The requirement of complete observations necessities data imputation when a moderate or large proportion of data is missing. However, inappropriate imputation would downgrade the performance of machine learning algorithms, leading to bad predictions and unreliable statistical inference. For high-dimensional large-scale mixed-type missing data, we develop a computationally efficient imputation method, missing value imputation via generalized factor models (MIG), under missing at random. The proposed MIG method allows missing variables to be of different types, including continuous, binary, and count variables, and are scalable to both data size n and variable dimension p while existing imputation methods rely on restrictive assumptions such as the same type of missing variables, the low dimensionality of variables, and a limited sample size. We explicitly show that the imputation error of the proposed MIG method diminishes to zero with the rate $O_p(\max\{n^{-1/2},~p^{-1/2}\})$ as both n and p tend to infinity. Five real datasets demonstrate the superior empirical performance of the proposed MIG method over existing methods that the average normalized absolute imputation error is reduced by 5.3%–34.1%.

Identification of Latent Subgroups for Time-varying Panel Data Models

♦*Ye He, Qing Luo, Liu Liu, Shengzhi Mao, Ling Zhou*

Sichuan Normal University, Sichuan Normal University, Chengdu University of Technology, Southwestern University of Finance and Economics, Southwestern University of Finance and Economics

In this paper, we propose a time-varying panel data model with latent group structures to simultaneously account for individual heterogeneity and smooth structural changes over time. To achieve computational efficiency without relying on prior group information, we introduce a novel centre-augmented K-harmonic means (KHM) procedure that encourages subjects to shrink toward their respective centers to identify latent group structures. We provide theoretical guarantees, including estimation consistency, accurate subgroup identification, and consistent selection of the number of groups. Numerical studies, including simulations and two real data examples, demonstrate the effectiveness of the proposed KHM method in accurately identifying the latent group structures in panel data.

A Functional Semiparametric Mixed Effects State Space Model with Prior Information for County Level Spatiotemporal Data

♦*Mengying You, Wensheng Guo*

Shanghai University of International Business and Economics, University of Pennsylvania

We propose a functional semiparametric mixed-effects state space model for analyzing COVID-19 case dynamics at the U.S. county level, incorporating both spatial structure and prior information. The model captures temporal trends using general smoothing splines and introduces spatially-structured random effects to borrow strength across geographically or epidemiologically similar counties. A state-space formulation enables efficient recursive estimation and forecasting, while accommodating both latent temporal processes and county-specific variability. The inclusion of prior knowledge on curve shapes enhances model stability in regions with sparse or noisy data. Through comparison with standard spline-based time series models, our approach demonstrates improved predictive accuracy, reduced forecast uncertainty, and greater capacity to recover shared epidemic patterns across counties. This framework is broadly applicable to spatiotemporal epidemiological surveillance and regional policy evaluation.

25CHI023: Contemporary survival analysis and new applications

SurGAN: A Generative Adversarial Network Model for Tabular Survival Data

♦*Hong Wang*

Central South University

Tabular data are the most commonly used form of medical data and are essential in survival analysis. Federated survival models which allow model training without sharing raw data have found increasingly applications in biomedical studies. However, federated learning framework often encounters non-independent and identically distributed data (Non-IID) problems across different data sources. In this study, we aim to mitigate the Non-IID problem in federated learning by augmenting tabular survival data via an improved generative adversarial network (GAN) model-SurGAN at the client level. Through a series of simulated and real data experiments, we evaluate the quality of the survival data generated by the SurGAN model and the performance of the federated survival model FSA-SurGAN. Experimental results have shown that the proposed deep generative model SurGAN has distinct advantages over existing generative models. Furthermore, in the federated learning setting with Non-IID data, the effectiveness of the FSA-SurGAN framework is also validated, demonstrating improved performance compared to standard federated and centralized survival models.

A model-free correlation coefficient for censored data

♦Linlin Dai, Tengfei Li, Kani Chen

Southwestern University of Finance and Economics, University of North Carolina at Chapel Hill, Hong Kong University of Science and Technology

This paper concerns assessing the association between covariates and right-censored outcomes. The challenge lies in the presence of partially unobservable time-to-event data, often due to study follow-up terminations. To tackle this issue, we first propose a simple and novel dependence measure that employs a known ``weighting" function. This pre-specified term eliminates the need for complex distribution approximations, thereby simplifying the construction of correlation coefficients for a broad range of data types, including right-censored data. The new measure has desirable properties: it ranges between 0 and 1, equaling 0 or 1 if and only if the two random variables are independent or one is a measurable function of the other. We then establish a consistent nonparametric correlation coefficient for right-censored data. It neither imposes constraints on the distributions of covariates and survival outcomes nor requires tuning parameters, and is also easy to calculate. Furthermore, it can effectively detect non-linear or non-monotonic relationships between covariates and the right-censored outcome, even under heavy censoring. To our knowledge, we are the first to devise a correlation coefficient for right-censored data with a limit satisfying all the mentioned desirable properties. We establish its asymptotic normality and also apply it to a permutation test for testing independence. Extensive empirical studies show its good performance in detecting various complicated dependencies.

Fiducial inference in survival analysis

♦Yifan Cui

Zhejiang University

In this talk, we introduce novel nonparametric and semiparametric fiducial approaches to censored survival data. We propose Gibbs samplers and establish Bernstein-von Mises theorems. We also demonstrate our estimators through extensive simulations and real data applications.

Nonparametric estimation of a state entry time distribution conditional on a (past) state occupation using current status data.

Samuel Anyaso-Samuel, ♦SOMNATH DATTA

NIH, U of Florida

Case-I interval-censored (current status) data from multistate systems are often encountered in biomedical and epidemiological studies. In this article, we focus on the problem of estimating state entry distribution and occupation probabilities, contingent on a preceding state occupation. This endeavor is particularly complex owing to the inherent challenge of the unavailability of directly observed counts of individuals at risk of transitioning from a state, due to severe interval censoring. We propose two nonparametric approaches, one using the fractional at-risk set approach recently adopted in the right-censoring framework and the other a new estimator based on the ratio of marginal state occupation probabilities. Both estimation approaches utilize innovative applications of concepts from the competing risks paradigm. The finite-sample behavior of the proposed estimators is studied via extensive simulation studies where we show that the estimators based on severely censored current status data have good performance when compared with those based on complete data. We demonstrate the application of the two methods to analyze data from patients diagnosed with breast cancer.

25CHI026: Efficient data collection and computing techniques in data-rich era

Maximum projection Latin hypercube designs using number theoretic methods

Yuxing Ye, Ru Yuan, ♦Yaping Wang

East China Normal University, Zhongnan University of Economics and Law, East China Normal University

Maximum projection (MaxPro) Latin hypercube designs (LHDs) are appealing for computer experiments where only a subset of the design factors are active. These designs are distinguished by their ability to optimize projection properties across all possible subsets of factors. However, the construction of MaxPro LHDs remains a challenging problem, and existing literature on the subject is limited. This paper presents two algebraic construction methods for MaxPro LHDs, based on good lattice point designs and number theory. The resulting designs are asymptotically optimal in terms of log-distance and nearly optimal in achieving the MaxPro lower bounds. Furthermore, they demonstrate excellent multi-criteria performance in terms of column orthogonality, maximin L1-distance, and the uniform projection criterion.

Data-driven Sampling Based Stochastic Gradient Descent Method

Yanjing Feng, Shiqi Zhou, ♦Yongdao Zhou

Nankai University, Nankai University, Nankai University

Sampling mechanism in mini-batch stochastic gradient descent (SGD) has been known to affect the convergence properties and the model performances. Despite the sustained efforts that make mini-batch SGD more data-efficient, it still remains a large area unexplored on how to select a batch of training data in each iteration. In this paper, we adopt a uniform design-based data-driven sampling method as batch selection technique, and based on the sampling results we construct one full gradient estimators to update model parameters. Theoretically we prove that without any convex assumption, the square L2-norm of full gradient with respect to the iterates generated by the combined method can converge at a sublinear rate in probability and the iteration complexity of the combined method is much lower than that for the vanilla mini-batch SGD. Futhermore, we develop a practical variant of proposed methods to save the computational cost. The numerical experiments empirically show that this practical variant outperforms other baseline methods and achieves a lower test error on real datasets in an efficient way.

A Wasserstein distance-based spectral clustering method for transaction data analysis

♦*Yingqiu Zhu, Danyang Huang, Bo Zhang*

University of International Business and Economics, Renmin University of China, Renmin University of China

With the rapid development of online payment platforms, it is now possible to record massive transaction data. Clustering on transaction data significantly contributes to analyzing merchants' behavior patterns. This enables payment platforms to provide differentiated services or implement risk management strategies. However, traditional methods exploit transactions by generating low-dimensional features, leading to inevitable information loss. In this study, we use the empirical cumulative distribution of transactions to characterize merchants. We adopt Wasserstein distance to measure the dissimilarity between any two merchants and propose the Wasserstein-distance-based spectral clustering (WSC) approach. Based on the similarities between merchants' transaction distributions, a graph of merchants is generated. Thus, we treat the clustering of merchants as a graph-cut problem and solve it under the framework of spectral clustering. To ensure feasibility of the proposed method on large-scale datasets with limited computational resources, we propose a subsampling method for WSC (SubWSC). The associated theoretical properties are investigated to verify the efficiency of the proposed approach. The simulations and empirical study demonstrate that the proposed method outperforms feature-based methods in finding behavior patterns of merchants.

BanditSIS: efficient algorithm for large-sample feature screening via multi-armed bandits

♦*Cheng Meng*

Renmin University of China

We consider the sure independence screening (SIS) method in Fan and Lv (2008), which is a standard feature screening approach that aims to eliminate non-informative features in high-dimensional datasets. The computational cost for SIS is at the order of O(np) for a predictor matrix of size n × p, thus may suffer from a huge computational burden when both n and p are considerable. Motivated by the multi-armed bandit (MAB) problem, we propose BanditSIS, a more efficient feature screening approach that reduces the computational cost to O(√np log(p) + n log(n)). The idea is to regard the marginal Pearson correlation as the reward, each feature as an arm, and the d important features as the best top d arms in the MAB problem. In such a framework, we aim to discard the features with the least expected marginal Pearson correlation with a certain level of accuracy and with a certain probability. Theoretically, we show our proposed method preserves the well-known sure screening property under mild regularity conditions. Numerical studies on various synthetic and real-world datasets demonstrate the superior performance of the proposed method in comparison with SIS, requiring significantly less computational time.

25CHI033: Innovations in Network Analysis

Temporal network analysis via a degree-corrected Cox model

Yuguo Chen, Lianqiang Qu, Jinfeng Xu, ♦*Ting Yan, Yunpeng Zhou*

University of Illinois at Urbana-Champaign, Central China Normal University, City University of Hong Kong, Central China Normal University, The University of Hong Kong

Temporal dynamics, characterised by time-varying degree heterogeneity and homophily effects, are often exhibited in many real-world networks.

As observed in an MIT Social Evolution study, the in-degree and out-degree of the nodes show considerable heterogeneity that varies with time. Concurrently, homophily effects, which explain why nodes with similar characteristics are more likely to connect with each other, are also time-dependent. To facilitate the exploration and understanding of these dynamics, we propose a novel degree-corrected Cox model for directed networks, where the way for degree-heterogeneity or homophily effects to change with time is left completely unspecified.

Because each node has individual-specific in- and out-degree parameters that vary over time, the number of the unknown time-dependent parameters in the model grows with the number of nodes. In other words, we are in a high-dimensional regime, and the estimation and inference of the unknown time-dependent parameters become nonstandard. We develop a local estimating equations approach to estimate the unknown parameters and establish the consistency and asymptotic normality of the proposed estimators in the high-dimensional regime. We further propose test statistics to check whether temporal variation or degree heterogeneity is present in the network. Simulation studies and a real data analysis are provided to assess the finite sample performance of the proposed method and illustrate its practical utility.

A community Hawkes model for continuous-time networks with interaction heterogeneity

Haosheng Shi, ♦*Wenlin Dai*

Renmin University of China, Renmin University of China

Continuous-time networks have attracted significant attention due to their widespread applications in various disciplines. A rich literature considers the community structure of the nodes, while few have accounted for the node heterogeneity of interaction propensities. To simultaneously account for both the self-exciting feature and the node heterogeneity, we propose a model based on the Hawkes process, which allows the interaction intensity to vary flexibly with incurred nodes and their affiliated communities. We derive the likelihood function using the immigration-birth representation of the Hawkes process and develop an innovative expectation-maximization algorithm with membership refinement to tackle the computational challenge. Further, we establish the consistency of parameter estimation under mild assumptions. The effectiveness of our model is validated by extensive simulation studies on synthetic data as well as two real-world applications.

Modeling reciprocity in directed networks

♦*Rui Feng, Chenlei Leng*

University of Warwick, University of Warwick

Reciprocity-the tendency for two individuals to form mutual connections-is a common feature of directed networks. This talk presents two models that incorporate covariates to capture reciprocity. The first introduces a novel Bernoulli model with reciprocity, along with an associated inference procedure. A key contribution is the analysis of effective sample sizes corresponding to different components of the model's parametrization. The second model extends the classical \$p_1\$ framework that incorporates node-specific heterogeneity by adding link-specific reciprocity. We develop a novel conditioning argument for estimating the reciprocity parameters and establish the minimax optimality of the resulting estimator. Numerical experiments are provided to support the theoretical results.

A Network Propagation Model for Graph-linked Data

♦*Yingying Ma, Chenlei Leng*

Beihang University, School of Economics and Management, University of Warwick

This study proposes a novel network propagation model that incorporates higher-order connected relationships to simultaneously capture the effects of both direct and indirect connections. The model can be viewed as an approximation of the linear-in-means framework. To estimate the unknown influence parameters, we introduce a naive least squares estimation approach and establish its consistency and asymptotic normality without requiring any distributional assumptions. Furthermore, we extend the model to accommodate nonlinear network propagation dynamics. The utility of both the linear and nonlinear models is demonstrated through simulations and real data analysis.

25CHI036: Innovative Approaches in Statistical Modeling and Analysis

Ratio-controlled screening for structural break predictive regressions

♦*Rongmao Zhang, Zhenjie Qin, Yang Zu*

Zhejiang Gongshang University, Zhejiang University, University of Macau

In this talk, we introduce a three-step efficient procedure to select and estimate the active predictors and change points of a structural break predictive regression, where the number of change points are allowed to vary with the sample size, and the predictors are allowed to be high-dimensional stationary, cointegrated and nonstationary, with sparse active variables. In this procedure, we first select the active predictors by a canonical correlation screening; then we estimate the change points by a ratio-controlled regression screening; and finally we eliminate the redundant break points and predictors by information criterion (IC). It is shown that the true break points and active predictors could be estimated and selected consistently. Simulations show that the proposed procedure performs quite well and is very efficient in computation.

Modeling paired binary data by a new bivariate Bernoulli model with flexible beta kernel correlation

Xunjian Li, Shuang Li, Guo-Liang Tian, ♦*Jianhua Shi*

Department of Statistics and Data Science, Southern University of Science and Technology, Department of Mathematics, Dongguan University of Technology, Department of Statistics and Data Science, Southern University of Science and Technology, School of Mathematics and Statistics, Minnan Normal University

Paired binary data often appear in studies of subjects with two sites such as eyes, ears, lungs, kidneys, feet and so on. This work aims to propose a new bivariate Bernoulli model with flexible beta kernel correlation for fitting the paired binary data with a wide range of group–specific disease probabilities. The correlation coefficient of the new model could be increasing, or decreasing, or unimodal, or convex with respect to the disease probability of one eye. To obtain the maximum likelihood estimates (MLEs) of parameters, we develop a series of minorization–maximization (MM) algorithms by constructing four surrogate functions with closed–form expressions at each iteration of the MM algorithms. Simulation studies are conducted, and two real datasets are analyzed to illustrate the proposed model and methods.

Flexible DNA Methylation Analysis scBS-Seq Data]{scFMA: A Flexible Random Effects Model for DNA Methylation Analysis with scBS-Seq Data

♦*Yanting Wu, Xifen Huang, Yao Lu, Jinfeng Xu, Hengjian Cui*

Yunnan Normal University

Background: Single-Cell Bisulfite Sequencing (scBS) has advanced rapidly because of its ability to study DNA methylation patterns at the level of individual cells. Since DNA methylation influences gene expression, a common approach involves identifying regions of differential methylation between cells to further explore transcriptional regulatory features and intercellular heterogeneity. However, this technique faces challenges such as limited sequencing depth and coverage, difficulties in processing large numbers of missing values, and the prevalence of extreme methylation rates (0 and 1) in single-cell data. These factors complicate accurate differential methylation analysis between cells.

Results: We develop a binomial model with flexible random effects for DNA methylation analysis. By combining the discrete approximation technique with the MM algorithm for model estimation and using the bootstrap method to construct the wald statistic for model testing, the proposed scFMA method identifies differentially methylated regions without restricting the form of the random-effects distribution, which solves the limitation that the existing methods are prone to model misspecification that leads to estimation bias.

Conclusion: Numerical studies have demonstrated that scFMA can identify more differentially methylated regions while maintaining low false positive rates across various distribution settings. We applied this method to investigate the mechanisms of cell methylation in early embryos. Despite the challenges posed by sparse data and small sample sizes, our model effectively captures the actual data distribution and reveals critical biological information during embryonic development.

Model free feature screening for large scale and ultrahigh dimensional survival data

Yingli Pan, Haoyu Wang, ♦*Zhan Liu*

Hubei University, Hubei University, Hubei University

A novel perspective on feature screening in the analysis of high dimensional right-censored large-p-large-N survival data is provided in this talk. The research introduces a distributed feature screening method known as Aggregated Distance Correlation Screening (ADCS). The proposed screening framework involves expressing the distance correlation measure as a function of multiple component parameters, each of which can be estimated in a distributed manner using a natural U-statistic from data segments. By aggregating the component estimates, a final correlation estimate is obtained, facilitating feature screening. Importantly, this approach does not necessitate any specific model specification for responses or predictors and is effective with heavy-tailed data. The study establishes the consistency of the proposed aggregated correlation estimator under mild conditions and demonstrates the sure screening property of the ADCS. Empirical results from both simulated and real datasets confirm the efficacy and practicality of the proposed ADCS approach.

25CHI038: Innovative methodology and strategy in statistical analysis

Application of Bayesian hierarchical model for subgroup analysis in vaccine efficacy study

Joyce Wang

Yingli Pan, Haoyu Wang, ♦Zhan Liu

Hubei University, Hubei University, Hubei UniversitySanofi

In traditional subgroup analyses, the treatment effects for each subgroup are typically estimated using only the data from that subgroup, without considering data from other subgroups within the same study. In contrast, a Bayesian hierarchical model assumes exchangeability of treatment effects across subgroups, implemented through random effect distributions. This approach allows information to be "borrowed" across subgroups, leading to more precise estimates of treatment effects for each subgroup than would be possible by relying solely on the data from that subgroup alone. In this session, methodology to apply this Bayesian hierarchical model in vaccine efficacy study will be discussed.

Missing data matter: an empirical evaluation of the impacts of missing EHR data in comparative effectiveness research

♦Yizhao Zhou, Jiasheng Shi, Ronen Stein, Xiaokang Liu, Robert Baldassano, Christopher Forrest, Yong Chen, Jing Huang

Department of Biometrics, China, Astrazeneca, Inc.

Objectives: The impacts of missing data in comparative effectiveness research (CER) using electronic health records (EHRs) may vary depending on the type and pattern of missing data. In this study, we aimed to quantify these impacts and compare the performance of different imputation methods. Materials and Methods: We conducted an empirical (simulation) study to quantify the bias and power loss in estimating treatment effects in CER using EHR data. We considered various missing scenarios and used the propensity scores to control for confounding. We compared the performance of the multiple imputation and spline smoothing methods to handle missing data.

Results: When missing data depended on the stochastic progression of disease and medical practice patterns, the spline smoothing method produced results that were close to those obtained when there were no missing data. Compared to multiple imputation, the spline smoothing generally performed similarly or better, with smaller estimation bias and less power loss. The multiple imputation can still reduce study bias and power loss in some restrictive scenarios, eg, when missing data did not depend on the stochastic process of disease progression. Discussion and Conclusion: Missing data in EHRs could lead to biased estimates of treatment effects and false negative findings in CER even after missing data were imputed. It is important to leverage the temporal information of disease trajectory to impute missing values when using EHRs as a data resource for CER and to consider the missing rate and the effect size when choosing an imputation method.

Integration of central statistical surveillance into central statistical surveillance (for binary endpoint)

♦Xiaojia Zhang

Sanofi

Central statistical surveillance plays a critical role in data monitoring for early detection of safety or efficacy signals in clinical trials. This topic explores the integration of Bayesian frameworks into central statistical surveillance for binary endpoints, addressing the need for adaptive, real-time decision-making under uncertainty. Leveraging Bayesian methods, we propose a dynamic surveillance system that updates posterior probabilities iteratively as new binary data (e.g., responder/non-responder) accumulate.

25CHI040: Innovative Statistical Learning Methods and Applications

Transformed dynamic quantile regression for case-cohort studies

♦Haijin He

Shenzhen University

In large cohort studies, the cost of collecting certain covariate information may be high. To reduce costs while maintaining comparable efficiency to cohort studies, case-cohort sampling designs are often used. However, the existing censored quantile regression models under the case-cohort design often require strict global linearity assumption or employ complex algorithms. In response to the two issues of the high research cost of queues and the excessively strict global linear assumption in the quantile regression model, this article proposes a class of power-transformed quantile regression models within the case-cohort framework. By introducing a process of power transformation, the proposed models have different transformation parameters at different quantile levels, relaxing the global linearity assumption and providing dynamic estimation of transformation parameters and covariate effects at various quantile levels. We introduce corresponding weight indicators based on the inverse probability weighting principle and presents a series of inverse probability weighted estimation equations for obtaining estimates of unknown parameters based on the zero-mean martingale processes. We provide an efficient algorithm based on minimizing L1-type convex functions and establishes the consistency and asymptotic normality of the proposed estimators through empirical process theory. We evaluate the finite-sample performance of the proposed methods through three numerical studies. Two real data sets are analyzed to illustrate the practical utility of our proposals.

Kernel Density Balancing with Application in Hi-C data

♦Ning Hao

The University of Arizona

High-throughput chromatin conformation capture (Hi-C) data provide insights into the 3D structure of chromosomes, with normalization being a crucial pre-processing step. A common technique for normalization is matrix balancing, which rescales rows and columns of a Hi-C matrix to equalize their sums. Despite its popularity and convenience, matrix balancing lacks statistical justification. In this talk, we introduce a statistical model to analyze matrix balancing methods and propose a kernel-based estimator that leverages spatial structure. Under mild assumptions, we demonstrate that the kernel-based method is consistent, converges faster, and is more robust to data sparsity compared to existing approaches.

Bayesian Inference of Phenotypic Plasticity of Cancer Cells Based on Dynamic Model for Temporal Cell Proportion Data

Shuli Chen, Yuman Wang, Da Zhou, ♦Jie Hu

Mounting evidence underscores the prevalent hierarchical organization of cancer tissues. At the foundation of this hierarchy reside cancer stem cells, a subset of cells endowed with the pivotal role of engendering the entire cancer tissue through cell differentiation. In recent times, substantial attention has been directed towards the phenomenon of cancer cell plasticity, where the dynamic interconversion between cancer stem cells and non-stem cancer cells has garnered significant interest. Since the task of detecting cancer cell plasticity from empirical data remains a formidable challenge, we propose a Bayesian statistical framework designed to infer phenotypic plasticity within cancer cells, utilizing temporal data on cancer stem cell proportions. Our approach is grounded in a stochastic model, adept at capturing the dynamic behaviors of cells. Leveraging Bayesian analysis, we scrutinize the moment equation governing cancer stem cell proportions, derived from the Kolmogorov forward equation of our stochastic model. Our methodology introduces an improved Euler method for parameter estimation within nonlinear ordinary differential equation models, also extending insights to compositional data. Extensive simulations robustly validate the efficacy of our proposed method. To further corroborate our findings, we apply our approach to analyze published data from SW620 colon cancer cell lines. Our results harmonize with in situ experiments, thereby reinforcing the utility of our method in discerning and quantifying phenotypic plasticity within cancer cells.

A new non-parametric resampling method based on representative points

♦Sirao Wang, Yinan Li, Kai-Tai Fang, Huajun Ye

As a resampling method, the bootstrap method is a powerful statistical tool that allows for estimating the distribution of a statistic (like the mean, variance, median, etc.) by resampling from the empirical distribution. This method consistently gains significant attention from the statistical community and is particularly useful when the underlying distribution of the data is unknown or when the sample size is too small for the assumptions of traditional parametric methods. In this paper, we propose a novel approach for constructing a non-parametric resampling method based on representative points. This resampling method is derived through the techniques of kernel smoothing and draws inspiration from the representative points. Theoretical analysis demonstrates that the convergence of resampling distribution constructed by the representative points can be guaranteed in the cases of the sample mean and sample variance under the Kolmogorov metric and Mallows-Wasserstein metric. To evaluate the efficiency of the novel method, a comprehensive Monte Carlo numerical study is conducted to compare this method with common non-parametric and parametric bootstrap methods. Simulation results show that it consistently improves the performance of non-parametric bootstrap methods in terms of the coverage rate of confidence intervals. Meanwhile, it is competitive in comparison with parametric bootstrap methods especially for small sample sizes.

25CHI042: Kernel methods in machine learning

On the Pinsker bound of inner product kernel regression in large dimensions

♦Weihao Lu, Jialin Ding, Haobo Zhang, Qian Lin

National University of Singapore, Tsinghua University, Tsinghua University, Tsinghua University

Building on recent studies of large-dimensional kernel regression, particularly those involving inner product kernels on the hypersphere $\mathcal{S}^{d}$, we investigate the Pinsker bound for inner product kernel regression in such settings. Specifically, we address the scenario where the sample size n is given by $\alpha d^{\gamma}(1+o_{d}(1))$ for some $\alpha, \gamma>0$. We have determined the exact minimax risk for kernel regression in this setting, not only identifying the minimax rate but also the exact constant, known as the Pinsker constant, associated with the excess risk.

Diffusion Actor-Critic: Formulating Constrained Policy Iteration as Diffusion Noise Regression for Offline Reinforcement Learning

♦Wenjia Wang

In offline reinforcement learning, it is necessary to manage out-of-distribution actions to prevent overestimation of value functions. One class of methods, policy-regularized methods, address this problem by constraining the target policy to stay close to the behavior policy. Although several approaches suggest representing the behavior policy as an expressive diffusion model to boost performance, it remains unclear how to regularize the target policy given a diffusion-modeled behavior sampler. In this paper, we propose Diffusion Actor-Critic (DAC) that formulates the Kullback-Leibler (KL) constraint policy iteration as a diffusion noise regression problem, enabling direct representation of target policies as diffusion models. Our approach follows the actor-critic learning paradigm that we alternatively train a diffusion-modeled target policy and a critic network. The actor training loss includes a soft Q-guidance term from the Q-gradient. The soft Q-guidance grounds on the theoretical solution of the KL constraint policy iteration, which prevents the learned policy from taking out-of-distribution actions. We demonstrate that such diffusion-based policy constraint, along with the coupling of the lower confidence bound of the Q-ensemble as value targets, not only preserves the multi-modality of target policies but also contributes to stable convergence and strong performance in DAC. Our approach is evaluated on the D4RL benchmarks and outperforms the state-of-the-art in nearly all environments.

Nonparametric Estimation of Mixed MNLs by Kernel Machine

♦Liang Ding

Fudan University

Non-parametric mixing multinomial logits are crucial in many classification problems given data are generated by mixture of unknown multinomial logits. In this work, we propose a non-parametric method based on tensors decomposition of kernel machines. We show that our model can achieves minimax convergence rate to the true mixture components as the data volume increases.

On non-redundant and linear operator-based nonlinear dimension reduction

Zhoufu Ye, ♦Wei Luo

Zhejiang University, Zhejiang University

Kernel principal component analysis (KPCA), a popular nonlinear dimension reduction technique, has the redundancy issue that each kernel principal component can be a measurable function of the preceding components. This harms the effectiveness of dimension reduction and leaves the dimension of the reduced data a heuristic choice. In this paper, we rebuild the theory of nonlinear dimension reduction centered on recovering the σ-field of the original data, and, using appropriate linear operators between RKHSs, we propose two sequential dimension reduction methods that address the redundancy issue, maintain the same level of computational complexity as KPCA, and rely on more plausible assumptions regarding the singularity of the original data. Compared with the existing nonlinear dimension reduction methods that also address the redundancy issue, our methods enjoy the parametric asymptotic rate and do not specify distributions on the reduced data, thereby preserving other patterns, if any, of the original data. By constructing a measure of the exhaustiveness of the reduced data, we also provide consistent order determination for these methods. Some numerical studies are presented at the end. The proposed work involves a novel characterization of conditional mean independence, which may attract independent research interest.

25CHI047: Modeling average and related topics

A Subsampling Strategy for AIC-based Model Averaging with Generalized Linear Models

Jun Yu, ♦HaiYing Wang, Mingyao Ai

Beijing Institute of Technology, University of Connecticut, Peking University

Subsampling is an effective approach to address computational challenges associated with massive datasets. However, existing subsampling methods do not consider model uncertainty. In this paper, we investigate the subsampling technique for the Akaike information criterion (AIC) and extend the subsampling method to the smoothed AIC model-averaging framework in the context of generalized linear models. By correcting the asymptotic bias of the maximized subsample objective function used to approximate the Kullback–Leibler divergence, we derive the form of the AIC based on the subsample. We then provide a subsampling strategy for the smoothed AIC model-averaging estimator and study the corresponding asymptotic properties of the loss and the resulting estimator. A practically implementable algorithm is developed, and its performance is evaluated through numerical experiments on both real and simulated datasets.

PEARL: Performance-enhanced Aggregated Representation Learning

♦Wenhui Li, Shijing Gong, Xinyu Zhang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, School of Management, University of Science and Technology of China, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Representation learning is a key technique in modern machine learning that helps models identify meaningful patterns in complex data. However, different methods extract different types of information, and relying on just one method may overlook important insights that could improve performance in downstream tasks. This paper proposes a performance-enhanced aggregated representation learning method, which combines multiple representation learning approaches to enhance downstream task performance. Our framework is designed to be general, accommodating a wide range of loss functions commonly used for evaluating machine learning models. To ensure computational efficiency, we use surrogate loss functions to make weight calculations more practical. Theoretically, we show that our method achieves optimal performance in downstream tasks, meaning that the risk of our estimator for the targeted task is asymptotically equivalent to the theoretical minimum. Additionally, we demonstrate that the proposed method effectively prioritizes models that can correctly approximate the targeted response, ensuring that the sum of weights assigned to correctly specified models converges to one when such models are included in the averaging pool. We evaluate our method against advanced machine learning models across different tasks, and the results show that it outperforms existing methods, making it a powerful tool for improving machine learning applications.

Functional sufficient dimension reduction with multivariate responses: A projection averaging method and beyond

♦Wenchao Xu

Shanghai University of International Business and Economics

We focus on the functional sufficient dimension reduction (FSDR) problem where the predictor is a function and the response is a vector. By projecting the multivariate response to univariate forms, we propose a projection averaging method for estimating the functional central subspace. The proposed method avoids multivariate smoothing of responses and can fully recover the central subspace under reasonable conditions. By employing functional sliced inverse regression, functional sliced average variance estimator, and functional cumulative slicing as the underlying univariate-response FSDR methods, we develop three specific methods. Additionally, we propose a class of distance-based methods. We establish the convergence rates for these methods under certain smoothness assumptions. Simulation studies and real data applications demonstrate the superior performance of the proposed methods.

Quantile Regression Model Averaging for Distributed Data

♦Haili Zhang

Shenzhen Polytechnic University

In the era of big data, the explosion of distributed data from diverse sources has reshaped the data analysis landscape, necessitating innovative methodologies for extracting meaningful insights. Quantile regression provides a nuanced understanding of the relationship between predictors and response variables at different points in the conditional distribution, rather than focusing solely on the mean.

This study investigates quantile regression model averaging in contexts characterized by large volumes of data that exhibit varying distributions across different regions or groups.

We propose a quantile regression model averaging estimator that effectively addresses the challenges posed by heterogeneous data sources, making it powerful for analyzing distributed data.

Our approach integrates the advantages of model selection and averaging techniques specifically designed to capture the various characteristics of distributed datasets. We prove the asymptotic optimality for the proposed model averaging estimator and demonstrate the effectiveness of the proposed model averaging estimator through comprehensive simulation studies and real-world applications, revealing its superiority over traditional methods. In addition, we tackle the computational challenges associated with distributed data processing and introduce a scalable algorithm that improves efficiency without compromising accuracy.

25CHI050: Modern Statistical Inference for Complex Data

Transfer Learning for Survival Data Using Pseudo Observations

♦Hanxuan Ye

University of Pennsylvania

Transfer Learning offers a powerful approach for leveraging data from related studies to enhance the target study outcomes. In survival analysis, understanding the relationship between survival probability and clinical/protein variables, possibly of high dimensional, can benefit significantly from integrating information across different studies. However, the challenges posed by censored survival data and distribution discrepancies among studies are substantial. This work addresses high-dimensional survival analysis within the context of transfer learning by employing pseudo-observations as labels. We design a transfer learning algorithm specifically tailored for censored survival data. The algorithm integrates the source and target data, which is an improvement over the single-task method that only uses the target data. Our approach is theoretically analyzed within a unified framework encompassing various models, including Cox and logistic models. Our method's efficacy is substantiated by extensive numerical studies and a real-world case study that is pertinent to cardiovascular disease, where it demonstrates superior performance.

Adaptive Independence Test via the Generalized HSIC

♦Yaowu Zhang

Shanghai University of Finance and Economics

Measuring and testing for nonlinear dependence between random vectors is a fundamental problem in the statistics and machine learning literature. Among various measures, the Hilbert-Schmidt independence criterion (HSIC) has garnered significant attention due to its theoretical and computational advantages. We analyze the traditional HSIC in depth and establish a connection between HSIC and the distance correlation. We reveal that it mainly detects linear dependencies and approaches zero when the dimensions grow, with the leading factor being the aggregation of linear dependencies. To improve the sensitivity of HSIC to nonlinear dependencies, we propose the Generalized HSIC (GHSIC), which has a closed form of expression and equals zero if and only if the two random vectors are independent. Through our investigation, we demonstrate that GHSIC effectively overcomes the limitations of HSIC and exhibits enhanced capability in detecting nonlinear dependencies, particularly in high-dimensional settings. Furthermore, we develop a data-adaptive test based on GHSIC, which outperforms the HSIC-based test in high-dimensional scenarios, even when linear dependencies are present. Extensive numerical experiments demonstrate the superiority of the proposed GHSIC.

Statistical Inference for Differentially Private Stochastic Gradient Descent

♦Zhanrui Cai, Xintao Xia, Linjun Zhang

The University of Hong Kong, Iowa State University, Rutgers University

Privacy preservation in machine learning, particularly through Differentially Private Stochastic Gradient Descent (DP-SGD), is critical for sensitive data analysis. However, existing statistical inference methods for SGD predominantly focus on cyclic subsampling, while DP-SGD requires randomized subsampling. This paper first bridges this gap by establishing the asymptotic properties of SGD under the randomized rule and extending these results to DP-SGD. For the output of DP-SGD, we show that the asymptotic variance decomposes into statistical, sampling, and privacy-induced components. Two methods are proposed for constructing valid confidence intervals: the plug-in method and the random scaling method. We also perform extensive numerical analysis, which shows that the proposed confidence intervals achieve nominal coverage rates while maintaining privacy.

Fast Association Recovery in High Dimensions by Parallel Learning

Ruipeng Dong, ♦Canhong Wen

University of Science and Technology of China, University of Science and Technology of China

Sparse reduced-rank regression is a widespread tool to reveal the association between multiple responses and predictors, and it has been widely applied to many data-driven applications. While much of the literature has studied related theoretical properties and numerical algorithms, due to high nonconvexity, the computation burden for large-scale datasets remains a great challenge in practice. Also, the gap between the statistical consistency and the algorithmic convergence needs more research. To address these two issues, we formulate a sparse reduced-rank regression as a set of parallel co-sparse unit-rank estimation problems and propose a new algorithm to estimate these subproblems in parallel. Under mild conditions, the iteration complexity of the proposed algorithm is polynomial with high-dimensional responses and predictors. We show a statistical consistency for the numerical solution, thereby bridging the gap between statistical consistency and numerical computation from nonconvex optimization. Moreover, the main calculation of the algorithm is restricted to a small active set, so it exhibits fast computation even in high dimensions. Extensive numerical studies and an application in genetics demonstrate the effectiveness and scalability of our approach.

25CHI052: Modern statistical methods in biostatistics

Self-Consistent Equation-guided Neural Networks for Censored Time-to-Event Data

Sehwan Kim, Rui Wang, ♦Wenbin Lu

Ewha Womans University, Department of Population Medicine,

Harvard Pilgrim Health Care Institute and Harvard Medical School, Department of Statistics, North Carolina State University

In survival analysis, estimating the conditional survival function given predictors is often of interest. There is a growing trend in the development of deep learning methods for analyzing censored time-to-event data, especially when dealing with high-dimensional predictors that are complexly interrelated. Many existing deep learning approaches for estimating the conditional survival functions extend the Cox regression models by replacing the linear function of predictor effects by a shallow feed-forward neural network while maintaining the proportional hazards assumption. Their implementation can be computationally intensive due to the use of the full dataset at each iteration because the use of batch data may distort the at-risk set of the partial likelihood function. To overcome these limitations, we propose a novel deep learning approach to non-parametric estimation of the conditional survival functions using the generative adversarial networks leveraging self-consistent equations. The proposed method is model-free and does not require any parametric assumptions on the structure of the conditional survival function. We establish the convergence rate of our proposed estimator of the conditional survival function. In addition, we evaluate the performance of the proposed method through simulation studies and demonstrate

its application on a real-world dataset.

A Multimodal Functional Deep Learning Approach for Multi-omics Data

Yuan Zhou, ♦Pei Geng, Shan Zhang, Feifei Xiao, Guoshuai Cai, Li Chen, Qing Lu

University of Florida, University of New Hampshire, Michigan State University, University of Florida, University of Florida, University of Florida, University of Florida

With rapidly evolving high-throughput technologies and consistently decreasing costs, collecting multimodal omics data in large-scale studies has become feasible. To tackle the challenge of complex interaction effects among omics levels, we propose a multimodal functional deep learning (MFDL) method for the analysis of high-dimensional multi-omics data. The MFDL method models the complex relationships between multi-omics variants and disease phenotypes through the hierarchical structure of deep neural networks and handles high-dimensional omics data using the functional data analysis technique. Furthermore, MFDL leverages the structure of the multimodal model to capture interactions between different types of omics data. Through simulation studies and real-data applications, we demonstrate the advantages of MFDL in terms of prediction accuracy and its robustness to the high dimensionality and noise within the data.

A new time-varying coefficients regression model for predicting COVID-19 deaths

♦Juxin Liu, Brandon Bellows, Joan Hu, Jianhong Wu, Zhou Zhou, Chris Soteros, Lin Wang

University of Saskatchewan, University of Saskatchewan, Simon Fraser University, York University, University of Toronto, University of Saskatchewan, University of New Brunswick

Since the beginning of the global pandemic of Coronavirus (SARS-COV-2), there have been many studies devoted to predicting COVID-19 related deaths. The aim of our work is to (1) explore the lagged dependence between the time series of case counts and the time series of death counts; and (2) utilize such a relationship for prediction. The proposed approach can also be applied to other infectious diseases or wherever dynamics in lagged dependence are of primary interest. Different from the previous studies, we focus on time-varying coefficient models to account for the evolution of the coronavirus. Using two different types of time-varying coefficient models, local polynomial regression models and piecewise linear regression models, we analyze the province-level data in Canada as well as country-level data using cumulative counts. We use out-of-sample prediction to evaluate the model performance. Our proposed methods can be easily and quickly implemented via existing R packages.

Assessing Algorithm Fairness Requires Adjustment for Risk Distribution Differences Across Population Subgroups: A Unified Framework for Fairness Evaluation

♦Xiaoyi Zheng, Hong Zhang, Sarah Hegarty, Jinbo Chen

Department of Statistics and Finance, University of Science and Technology of China, Anhui, China, Department of Statistics and Finance, University of Science and Technology of China, Anhui, China, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

The proliferation of risk prediction models and algorithm-driven decision making has led to increased demand for methods to detect, assess, and mitigate algorithmic bias. Existing fairness assessment frameworks typically evaluate equality by assessing the consistency of performance metrics, such as TPR, across protected subgroups. However, we highlight a critical but overlooked issue: these performance metrics may vary when the proportion of individuals with the same underlying risk, that is, the risk distribution, varies across these subgroups, even if the model equally captures the underlying risks. Failure to account for variations in risk distributions may lead to misleading conclusions about performance disparity across subgroups. To address this issue, we propose a unified framework for fairness assessment that adjusts existing fairness metrics to account for subgroup-specific risk distributions. We also develop statistical procedures for parameter estimation and inference under this framework, and demonstrate the effectiveness of our method through applications to both simulated and real-world data.

25CHI064: Recent Advance in High-dimensional Modelling

Testing High-Dimensional Effects in Quantile Regression with High-Dimensional Confounding: A Decorrelated Smoothing Approach

♦Peirong Xu

Shanghai Jiao Tong University

Quantile regression provides a flexible framework in econometrics for modeling the relationship between a response variable and multivariate predictors, particularly when heterogeneous effects are present. This paper addresses the challenge of testing high-dimensional coefficients in quantile regression in the presence of high-dimensional nuisance parameters. We first extend a recently developed score test by utilizing a convolution-type smoothed quantile loss, which facilitates efficient computation and scalable inference. We derive the asymptotic distributions of the proposed test under both the null and local alternatives, assuming weak correlations between covariates of interest and nuisance covariates. To mitigate the bias caused by strong correlations, we further propose a decorrelated smoothing score test, which improves Type-I error control and power performance. Theoretical properties, including the limiting null distribution and power behavior, are established in settings where both the target and nuisance parameters are ultra-high dimensional. Comprehensive simulation results validate our theory and demonstrate the robustness and effectiveness of the proposed test when dealing with heavy-tailed and asymmetric data. We further illustrate the practical utility of our method through an empirical analysis of U.S. stock market data.

The Inferential Theory of Random Projection Methods

♦Hongjun Li

Tsinghua University

Random projection methods are widely employed for dimensionality reduction in statistical and machine learning applications, preserving essential geometric and structural properties of high-dimensional data. This work investigates how such projections influence statistical inference, with a focus on balancing computational efficiency and statistical validity. By leveraging foundational results like the Johnson-Lindenstrauss (JL) Lemma for geometric preservation and the Restricted Isometry Property (RIP) for signal recovery, we formalize the theoretical justification for dimensionality reduction in inferential tasks. We address key challenges—including bias correction, variance inflation, and interpretability—that arise from the randomness introduced by projections. To operationalize this framework, we develop an estimator that integrates random projections and establish its asymptotic properties, such as consistency and asymptotic normality. Finally, we validate our methodology through Monte Carlo simulations and empirical case studies, demonstrating the estimator's practical performance and the robustness of its inferential guarantees in real-world settings.

Low-rank and sparse network regression

Aureo de Paula, ♦Yingxing Li, Weining Wang

University College London, Xiamen University, University of Bristol.

This paper analyzes spillover effects in spatial network models under settings where measurement noises might contaminate the neighborhood (i.e. adjacency) matrix. We propose to adopt the low-rank and sparse structure to capture the stylized network pattern in empirical datasets. We develop a robust estimation framework via regularization techniques: the Least Absolute Shrinkage and Selection Operator (LASSO) for the sparse component and a nuclear norm penalty for the low-rank component. We propose two estimators: (1) A two-stage procedure that first de-noises the adjacency matrix via regularization and subsequently integrates the purified network to regression analysis, and (2) A single-step supervised Generalized Method of Moments (GMM) estimator jointly estimates the regression parameters and refines the network structure. Simulation evidence underscores the superiority of our approach. In scenarios with noisy net-works, our method reduces the root mean squared error (RMSE) of coefficient estimates by 50–70% compared to conventional GMM. This advantage is more significant when network contamination is endogenous, which is a common challenge in empirical settings where measurement errors are correlated with the observed outcomes. Applied to the dataset of Besley and Case (1995), our framework demonstrates practical utility. The decomposition not only improves estimation relia- bility but also generates granular insights for policy design. These results highlight how explicitly modeling network structure heterogeneity can bridge methodological rigor and policy relevance.

Estimation of Large Dynamic Precision Matrices with a Latent Semiparametric Structure

♦Jia Chen, Yuning Li, Oliver Linton

University of Macau, University of York, University of Cambridge

This paper studies the estimation of dynamic precision matrices with multiple conditioning variables for high-dimensional time series. We assume that the high-dimensional time series has an approximate factor structure plus an idiosyncratic error term, allowing the time series to have a non-sparse dynamic precision matrix and hence, enhancing the applicability of our method. Using the Sherman-Morrison-Woodbury formula, the estimation of the dynamic precision matrix for the time series boils down to the estimation of a low-rank factor structure and the precision matrix of the idiosyncratic error term. For the latter, we introduce an easy-to-implement semiparametric method to estimate the entries of the corresponding dynamic covariance matrix via the Model Averaging MArginal Regression (MAMAR) before applying the constrained l_1 minimisation for inverse matrix estimation (CLIME) method to obtain the dynamic precision matrix. Under some regularity conditions, we derive the uniform consistency for the proposed estimators. We provide a simulation study that illustrates the finite-sample performance of the developed methodology and an application in construction of minimum variance portfolios using daily returns of S&P 500 constituents from 2000 to 2023.

25CHI091: semi-parametric and nonparametric methods for complex data analysis

Statistical methods for transfer learning in survival analysis

♦Yu Gu, Donglin Zeng, Danyu Lin

University of Hong Kong, University of Michigan, University of North Carolina at Chapel Hill

Transfer learning has gained considerable interest in computer science and statistics in recent years. Traditional transfer learning methods typically assume similar distributions between target and source populations. However, this distributional similarity assumption is often violated under semiparametric survival models. In this talk, I will present novel transfer learning methods for survival analysis. Our methods rely on less stringent assumptions and achieve better prediction performance than existing methods in both numerical and real-data studies. When the source information is sufficiently accurate, our estimator enjoys a faster convergence rate than the target-only estimator.

Spatial deconvolution and cell type-specific spatially variable gene detection in spatial transcriptomics

♦*Yuehua Cui*

Michigan State University

Spatial transcriptomics (ST) provides crucial insights into tissue-specific gene expression patterns in various cancer studies. Most ST data, such as those obtained from the 10x Visium platform, are captured at a spot resolution which measures gene expression across multiple cells, often originating from various cell types. Deconvolution of such multi-cellular data to infer cell type compositions is crucial for further downstream analysis. Recent methodological developments have greatly advanced the detection of spatially variable genes (SVGs), whose expression patterns are non-random across tissue locations. Given that many SVGs correlate with cell type compositions, we introduce a unified approach to identify both SVGs and cell type-specific SVGs (ctSVGs), integrated with ST deconvolution, under a linear mixed-effect model framework. Our method, termed STANCE, ensures tissue rotation-invariant results, with a two-stage testing strategy: initial SVG/ctSVG detection followed by ctSVG-specific testing. We demonstrate its performance through extensive simulations and analyses of public datasets. Downstream analyses reveal STANCE's potential in spatial transcriptomics analysis.

Checking the Cox Proportional Hazards Model with Interval-Censored Data

♦*Yangjianchen Xu, Donglin Zeng, Danyu Lin*

University of Waterloo, University of Michigan, University of North Carolina at Chapel Hill

We present a general framework for checking the adequacy of the Cox proportional hazards model with interval-censored data, which arise when the event of interest is known only to occur over a random time interval. Specifically, we construct certain stochastic processes that are informative about various aspects of the model, i.e., the functional forms of covariates, the exponential link function and the proportional hazards assumption. We establish their weak convergence to zero-mean Gaussian processes under the assumed model through empirical process theory. We then approximate the limiting distributions by Monte Carlo simulation and develop graphical and numerical procedures to check model assumptions and improve goodness of fit. We evaluate the performance of the proposed methods through extensive simulation studies and provide an application to the Atherosclerosis Risk in Communities Study.

Distributional Off-Policy Evaluation with Deep Quantile Process Regression

♦*Fan Zhou*

Shanghai University of Finance and Economics

This paper investigates the off-policy evaluation (OPE) problem from a distribu- tional perspective, with the aim of modeling the entire distribution of total returns, rather than focusing solely on estimating the expectation (value function), as most existing OPE methods do. Specifically, we introduce a quantile-based approach for OPE using deep quantile process regression, presenting a novel algorithm called Deep Quantile Process regression-based Off-Policy Evaluation (DQPOPE). We provide new theoretical insights into the deep quantile process regression technique, extending ex- isting approaches that estimate discrete quantiles to estimate a continuous quantile function. A key contribution of our work is the rigorous sample complexity analysis for distributional OPE with deep neural networks, bridging theoretical analysis with practical algorithmic implementations. We show that DQPOPE achieves statistical efficiency by estimating the full return distribution using the same sample size re- quired to estimate a single policy value using conventional methods. Furthermore, our empirical studies illustrate that DQPOPE provides significantly more precise and robust policy value estimates than standard methods, thereby enhancing the practical applicability and effectiveness of distributional reinforcement learning approaches.

25CHI070: Recent advances in network modeling

Moment-integrated Bias-adjusted Spectral Method for Community Detection in Multi-layer Networks

♦*Xuefei Wang, Junhui Wang, Gaorong Li*

School of Statistics, Beijing Normal University, Department of Statistics, The Chinese University of Hong Kong, School of Statistics, Beijing Normal University

Multi-layer networks not only directly provide adjacency matrices but also indicate between-layer correlation. It is of vital importance for accurately detecting communities to effectively capture information from the network data. In this paper, under the framework of multi-layer stochastic block model, a Spectral method with Moments integration and Bias Adjustment (SpecMBA) is provided for community detection. The key distinguishing feature of SpecMBA is that it complements insufficient individual-layer information through balancing the first and second moments of layer-wise adjacency matrices with a hyperparameter α. In addition, a data-driven likelihood-based approach is proposed to choose the optimal α. Therefore, the flexibility in adapting to network data makes SpecMBA stand out prominently, which has been confirmed in the numerical study. Under mild conditions, especially weak sparsity restriction, the community detection consistency for SpecMBA is established. Additionally, the application on the international food trading network reveals interesting findings.

Data Integration: Network-Guided Covariate Selection in High-Dimensional Data

♦*Wanjie Wang, Tao Shen*

National University of Singapore, National University of Singapore

When integrating datasets from different studies, it is common that they have components of different formats. How to combine them organically for improved estimation is important and challenging. This paper investigates this problem in a two-study scenario, where covariates are observed for all subjects, but network data is available in only one study, and response variables are available only in the other.

To leverage the partially observed network information, we propose the Network- Guided Covariate Selection (NGCS) algorithm. It integrates the spectral information from network adjacency matrices with the Higher Criticism Thresholding approach for informative covariates identification. Theoretically, we prove that NGCS achieves the optimal rate in covariate selection, which is the same rate in the supervised learning setting. Furthermore, this optimality is robust to network models and tuning parameters.

This framework extends naturally to clustering and regression tasks, with two proposed algorithms: NG-clu and NG-reg. For clustering, NG-clu accurately clus- ters data points despite incomplete network information. For regression, NG-reg enhances predictive performance by incorporating latent covariate structures inferred from network data. Empirical studies on synthetic and real-world datasets demon- strate the robustness and superior performance of our algorithms, underscoring their effectiveness in handling heterogeneous data formats.

A dynamic network autoregressive model for time-varying network-link data

♦*Jingnan Zhang, Bo Zhang, Yu Chen*

University of Science and Technology of China, University of Science and Technology of China, University of Science and Technology of China

Network-linked data, where different units are linked through a network has been extensively studied in literature. However, its extension, specifically time-varying network-link data, has received less investigation. Existing methods for time-varying network-link data only assume that units' attributes change over time, neglecting network evolution. To address this gap, we propose a dynamic network autoregressive model for time-varying network-link data, where both units' attributes and networks are allowed to vary over time. A tensor decomposition method is employed to provide low-dimensional embedding vectors, which are further used to reformulate the traditional network autoregressive model. Interestingly, node-embedding vectors are concentrated around some group centers but are not exactly the same within some groups. Meanwhile, both within-group and global homogeneities are considered for the effect of covariate vectors. To tackle the resultant optimization task, we develop the power update algorithm and an efficient alternative updating algorithm. Furthermore, the asymptotic consistencies of the proposed method are established, irrespective of the presence of the global effect of covariate vector. These consistencies are demonstrated by extensive simulated examples and a real example of time-varying network-linked fund data.

False Discovery Rate Control Using Bi-Gaussian Mirrors

♦*Binyan Jiang*

The Hong Kong Polytechnic University

Effectively controlling the false discovery rate (FDR) in high-dimensional variable selection is a fundamental statistical problem that has garnered significant research interest. In this paper, we propose a novel, user-friendly, and computationally efficient method called Bi-Gaussian Mirrors (BGM), which offers a conceptually simple yet powerful approach to FDR control. Our method makes the first attempt to achieve FDR control in high-dimensional data with complex dependencies, while overcoming key limitations of existing approaches, such as prior knowledge of the joint distribution of data, significant power loss, the need for full symmetry in test statistics, and the theoretical restriction to linear regression models. Additionally, we present a self-guiding procedure designed to enhance the practicality and applicability of the BGM method. Theoretical guarantees for FDR control and asymptotic power are rigorously established under regularity conditions. Moreover, extensive numerical simulations and two real-data examples demonstrate that the BGM method outperforms existing approaches in terms of finite-sample performance, achieving a superior balance between FDR control and testing power.

25CHI080: Recent developments in analyzing complex data

A Unified Analysis of Likelihood-based Estimators in the Plackett-Luce Model

♦*Ruijian Han, Yiming Xu*

The Hong Kong Polytechnic University, University of Kentucky

The Plackett–Luce model has been extensively used for rank aggregation in social choice theory. A central question in this model concerns estimating the utility vector that governs the model's likelihood. In this paper, we investigate the asymptotic theory of utility vector estimation by maximizing different types of likelihood, such as full, marginal, and quasi-likelihood. Starting from interpreting the estimating equations of these estimators to gain some initial insights, we analyze their asymptotic behavior as the number of compared objects increases. In particular, we establish both the uniform consistency and asymptotic normality of these estimators and discuss the trade-off between statistical efficiency and computational complexity. For generality, our results are proven for deterministic graph sequences under appropriate graph topology conditions. These conditions are shown to be revealing and sharp when applied to common sampling scenarios, such as nonuniform random hypergraph models and hypergraph stochastic block models. Numerical results are provided to support our findings. This is a joint work with Yiming Xu.

Approximation Error from Discretizations and Its Applications

♦*Junlong Zhao, Xiumin Liu, Bin Du, Yufeng Liu*

Beijing Normal University, Beijing Normal University, Beijing Normal University, University of North Carolina at Chapel Hill

Converting a continuous variable into a discrete one is a commonly used technique for various problems in both statistics and machine learning. It is well known that discretizations result in biases. However, this issue has not been studied systematically. In this paper, a general framework is proposed to understand and compare the approximation errors of different slicing strategies. Poincar´e-type inequalities are first established for univariate discretizations and then generalized to the multivariate and other settings. It is shown that the bias is controlled by two factors: the distance between two specific distributions that are generated with and without discretizations respectively, and the smoothness of the functions involved. Several important applications are considered to illustrate the usefulness of the results. In particular, our results help to answer some open problems in the literature of dimension reduction. Furthermore, as an illustration of the usefulness of discretizations, we propose an algorithm for regression problems, by combining random forest with partial discretizations of responses. Simulation results confirm the advantages of this algorithm over the classical random forest.

Large-Scale Curve Time Series with Common Stochastic Trends

♦*Degui Li, Yuning Li, Peter C.B. Phillips*

University of Macau, University of York, Yale University

In this paper, we study high-dimensional curve time series with common stochastic trends. We adopt a dual functional factor model structure with a high-dimensional factor model for the observed curve time series and a low-dimensional factor model for the latent curves with common trends. A functional PCA technique is applied to estimate the common stochastic trends and functional factor loadings. Under some regularity conditions, we derive the mean square convergence and limit distribution theory for the developed estimates, allowing the dimension and sample size to jointly diverge to infinity. We also propose an easy-to-implement criterion to consistently select the number of common stochastic trends and further discuss the model estimation when the nonstationary factors are cointegrated. Extensive Monte-Carlo simulation studies and two empirical applications to large-scale temperature curves in Australia and log-price curves of the S&P stocks, respectively, are conducted to illustrate the finite-sample performance of the developed methodology.

Supervised Factor Modeling for High-Dimensional Linear Time Series

♦Guodong Li

University of Hong Kong

Motivated by Tucker tensor decomposition, this paper imposes low-rank structures to the column and row spaces of coefficient matrices in a multivariate infinite-order vector autoregression (VAR), which leads to a supervised factor model with two factor modelings being conducted to responses and predictors simultaneously. Interestingly, the stationarity condition implies an intrinsic weak group sparsity mechanism of infinite-order VAR, and hence a rank-constrained group Lasso estimation is considered for high-dimensional linear time series. Its non-asymptotic properties are discussed by balancing the estimation, approximation and truncation errors. Moreover, an alternating gradient descent algorithm with hard-thresholding is designed to search for high-dimensional estimates, and its theoretical justifications, including statistical and convergence analysis, are also provided. Theoretical and computational properties of the proposed methodology are verified by simulation experiments, and the advantages over existing methods are demonstrated by analyzing US quarterly macroeconomic variables.

25CHI087: Recent Statistical Advances in Complex Genetic and Genomic Data Analysis

Differential Inference for Single-cell RNA-Sequencing Data

Fangda Song, Kevin Yip, ♦Yingying Wei

The Chinese University of Hong Kong, Shenzhen, Sanford Burnham Prebys Medical Discovery Institute, The Chinese University of Hong Kong

Single-cell RNA-sequencing (scRNA-seq) experiments are becoming increasingly complicated with multiple treatment or biological conditions. However, guidelines on experimental designs and rigorous statistical methods for comparative scRNA-seq studies with cells collected from multiple conditions are still lacking. For a confounded design, the batch effects, cell-type effects and condition effects can never be distinguished. Therefore, we mathematically derive the requirements for a valid design for a comparative scRNA-seq study. Moreover, existing methods for identifying differentially expressed genes and differential cell-type abundance between conditions have to be multi-stage approaches. Because multi-stage approaches ignore uncertainties in previous stages and may propagate errors from earlier stages to later stages, they can suffer from high error rates. Here, we introduce DIFseq, a hierarchical model that accounts for all uncertainties and hence rigorously quantifies the condition effects on both cellular composition and cell-type-specific gene expression levels. DIFseq substantially outperforms state-of-the-art methods for both simulated and real data.

Genetic association testing with multivariate survival phenotypes under interval censoring

Juhee Lee, ♦Chenxi Li, Gongjun Xu, Qing Lu

Michigan State University, Michigan State University, University fo Michigan, University of Florida

Kernel-based multi-marker tests have been popular in association studies with genotyped or sequencing data due to their ability to improve power and reduce the burden of multiple testing compared to likelihood-based joint tests and single-variant tests. Many kernel-based multi-marker tests have been developed for various types of outcomes, including continuous, categorical and survival phenotypes. However, applying existing multi-marker survival tests to study the genetic architecture underlying early childhood caries is challenging due to two complexities: the correlation between caries onset times across multiple teeth and the interval censoring of those times.

We develop two solutions to this problem:

1) a set of two weighted V statistic-based multi-marker tests for multivariate interval censored data, and 2) a p-value combination approach based on an existing multi-marker test for univariate interval-censored survival phenotypes. Our simulation studies show that the three methods can control Type I error rates and have decent power under modest sample sizes, and no single method outperforms the others in terms of power in all the simulation scenarios. An application to the ZOE 2.0 study illustrates the practical utility of the proposed methods.

A unified framework for identification of cell-type-specific spatially variable genes in spatial transcriptomic studies

Zhiwei Wang, Yeqin Zeng, Ziyue Tan, Yuheng Chen, Xinrui Huang, Hongyu Zhao, Zhixiang Lin, ♦Can Yang

HKUST, HKUST, HKUST, HKUST, HKUST, Yale, CUHK, HKUST

Characterizing spatial variations in gene expression with cell type specificity is crucial for understanding complex diseases, yet it remains challenging. Here, we introduce the Mixture of Mixed Models (MMM), a unified framework that effectively identifies cell-type-specific spatially variable genes (SVGs) in spatial transcriptomic studies. MMM achieves robust performances because of our innovations in model and algorithm design. In the mouse brain study, MMM reveals that these SVGs are significantly enriched for heritability in brain-related phenotypes, highlighting their importance in complex traits and diseases. Our findings also shed light on the role of cell-type-specific SVGs in cell-cell communications and microenvironment regulation, advancing our understanding of complex tissues.

Hypothesis testing in high-dimensional censored-

transformation models

Xiao Zhang, Xiangyong Tan, Runze Li, ♦Xu Liu

The Chinese University of Hong Kong, Shenzhen, Jiangxi University of Finance and Economics, Pennsylvania State University, Shanghai University of Finance and Economics

With the rapid development of modern technologies, high-dimensional statistical inference of survival times has become increasingly important in various fields, including biostatistics and financial risk. Herein, we propose an efficient rank-based test statistic that is asymptotically normally distributed. The proposed test statistic allows for covariance-dependent censoring and is robust against heavy-tailed distributions and potential outliers. Furthermore, for practical purposes, we propose a new rank-based test statistic to test for the existence of high-dimensional features with high-dimensional control factors. Asymptotic distributions under the null hypothesis and local alternatives are established for the proposed test statistic. Numerical studies are performed to evaluate the finite-sample performance of the proposed test statistics. We illustrate the proposed method to an empirical analysis of a skin cutaneous melanoma (SKCM) dataset.

25CHI089: Sample Size, Power, and Likelihood

Influence Function-based Empirical Likelihood for AUC in Presence of Covariates

Baoying YANG, Xinjie HU, ♦Gengsheng Qin

Southwest Jiaotong University, CHINA, Georgia State University, USA, Georgia State University, USA

In ROC analysis, the area under the ROC curve (AUC) is a popular one number summary of the discriminatory accuracy of a diagnostic test. AUC measures the overall diagnostic accuracy of a test but fails to account for the effect of covariates when covariates are present and associated with the test results. Adjustment for covariate effects can greatly improve the diagnostic accuracy of a test. In this paper, using information provided by the influence function, empirical likelihood methods are proposed for inferences of AUC in presence of covariates. For parameters in the AUC regression model, it is shown that the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic is a standard chi-square distribution. Hence, confidence regions for the regression parameters can be obtained without any variance estimation. Simulation studies are conducted to compare the finite sample performances of the proposed Empirical Likelihood (EL) based methods with the existing Normal Approximation (NA) based method in the AUC regression. Simulation results indicate that the Bootstrap-calibrated Influence Function-based Empirical Likelihood (BIFEL) confidence region outperforms the NA-based confidence region in terms of coverage probability. We also propose an interval estimation method for the covariate-adjusted {AUC} based on the BIFEL confidence region. Finally, we illustrate the recommended method with a real prostate-specific antigen (PSA) data example.

Optimizing Sample Size in vaccine efficacy trial: integrating the timing expectation of interim analysis and the seasonallity prediction of diseases

♦Penny Peng

Department of Biostatistics and Programming, China, Sanofi, Inc.

This study presents a novel approach to sample size optimization in vaccine efficacy trials that incorporates both the timing of interim analyses and seasonal disease incidence patterns. The proposed approach aims to ensure that the target number of cases is achieved within the predefined median follow-up period. This strategy minimizes the risk of prolonged follow-up, during which waning efficacy may lead the trail failure. By simulating various enrollment scenarios and disease incidence patterns, we demonstrate how strategic timing of interim analyses can significantly improve trial efficiency while maintaining statistical rigor. This approach is particularly valuable for infectious diseases with known seasonal variations, enabling more efficient vaccine development pathways without compromising scientific validity.

The Identifiability of Copula Models for Dependent Competing Risks Data With Exponentially Distributed Margins

♦Antai Wang

New Jersey Institute of Technology

We prove the identiability property of Archimedean copula models for dependent competing risks data when at least one of the failure times is ex-ponentially distributed. With this property, it becomes possible to quantify the dependence between competing events based on exponentially distributed dependent censored data. We demonstrate our estimation procedure using simulation studies and in an application to survival data.

Considering the correlation between serotypes for the sample size estimation in vaccine clinical trials

♦Jieqi Jin, Fabrice Bailleux, Jian Ding, Ian Deng

Department of Biostatistics and Programming, China, Sanofi, Inc.

In literature of previous study, it was found that considering the correlation between serotypes when calculating the sample size can avoid underestimation of power and thus reduce the sample size of trial. We took a study as an example to introduce the way we considered the correlation when calculating the sample size, the macro we used with the information of parameters, the results and discussion.

25CHI092: Some important topics in pharmaceutical statistics

Multiple Comparisons Procedures for Analyses of Joint Primary Endpoints and Secondary Endpoints

Xiaolong Luo, Lerong Li, Oleksandr Savenkov, Weijian Liu, Xiao Ni, Weihua Tang, ♦Wenge Guo

Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, Sarepta Therapeutics, New Jersey Institute of Technology

One of the main challenges in drug development for rare diseases is selecting the appropriate primary endpoints for pivotal studies. Although many endpoints can effectively reflect clinical benefit, their sensitivity often varies, making it difficult to determine the required sample size for study design and to interpret final results, which may be underpowered for some or all endpoints. This complexity is further compounded when there is a desire to support regulatory claims for multiple clinical endpoints and dose regimens due to the issues of multiplicity and sample size constraints. Joint Primary Endpoints (JPEs) offer a compelling strategy to address these challenges; however, their analysis in conjunction with component endpoints presents additional complexities, particularly in managing multiplicity concerns for regulatory claims. To address these issues, this paper introduces a robust two- stage gatekeeping framework designed to test two hierarchically ordered families of hypotheses. A novel truncated closed testing procedure is employed in the first stage, enhancing flexibility and adaptability in the evaluation of primary endpoints. This approach strategically propagates a controlled fraction of the error rate to the second stage for assessing secondary endpoints, ensuring rigorous control of the global family- wise Type I error rate across both stages. Through extensive numerical simulations and real- world clinical trial applications, we demonstrate the efficiency, adaptability, and practical utility of this approach in advancing drug development for rare diseases while meeting stringent regulatory requirements.

Matching-Assisted Power Prior for Incorporating Real-World Data in Clinical Trials

Ruoyuan Qian, Biqing Yang, Xinyi Xu, ♦Bo Lu

The Ohio State University, The Ohio State University, The Ohio State University, The Ohio State University

Leveraging external data information to supplement randomized clinical trials has been a popular topic in recent years, especially for medical device and drug discovery. In rare diseases, it is very challenging to recruit patients and run a large-scale randomized trial. To take advantage of real-world data from historical trials on the same disease, we can run a small hybrid trial and borrow historical controls to increase the power. But the borrowing needs to be conducted in a statistically principled manner. Bayesian power prior methods and propensity score adjustments have been discussed in the literature. In this paper, we propose a matching-assisted power prior approach to better mitigate overt bias. A subset of comparable external controls is selected by groups through template matching, and different weights are assigned to these groups based on their similarity to the current study population. Power priors are then implemented to incorporate the information into Bayesian inference. Unlike conventional power prior methods, which discount all controls similarly, matching pre-selects good controls, hence improves the quality of external data being borrowed. We compare its performance with the existing propensity score-integrated power prior approach through simulation studies and illustrate the implementation using data from a real acupuncture clinical trial.

Consistency consideration for a region in a multiple regional clinical trial

♦En-Tzu (Angela) Tang

Abbvie Inc.

This presentation will explore key considerations for planning sample size of a region and evaluating consistency with overall results in multiregional clinical trials (MRCTs). Factors such as trial design, endpoint type, and different participation models for patients of a specific region (eg, MRCT, extension cohorts, or standalone studies) will be discussed, along with general strategies to support regulatory acceptance.

Sequential Monitoring of Covariate Adaptive Randomized Clinical Trials with Nonparametric Approaches

Xiaotian Chen, Jun Yu, ♦Hongjian Zhu, Li Wang

AbbVie Inc., AbbVie Inc., Systlmmune Inc., AbbVie Inc.

The importance of covariate adjustment in clinical trials has been underscored by the U.S. FDA's guidance. Inference, with or without covariates, after implementing covariate adaptive randomization (CAR), is garnering increased interest. This talk investigates the sequential monitoring of covariate-adaptive randomized clinical trials through nonparametric methods, a critical advancement for enhancing the precision and efficiency of medical research. CAR, which incorporates baseline patient characteristics into the randomization process, aims to mitigate the risk of confounding and improve the balance of covariates across treatment groups, thereby addressing patients' heterogeneity. Although CAR is known for its benefits in reducing biases and enhancing statistical power, its integration into sequentially monitored clinical trials—a standard practice—poses methodological challenges, particularly in controlling the type I error rate. By employing a nonparametric approach, we demonstrate through theoretical proofs and numerical analyses that our methods effectively control the type I error rate and surpass traditional randomization and analysis methods. This study not only fills a gap in the literature on sequential monitoring of CAR without model misspecification but also proposes practical solutions for enhancing trial design and analysis, thereby contributing significantly to the field of clinical research.

25CHI094: Specific Statistical considerations in clinical trial design

Statistical consideration of estimands in vaccine clinical trials

♦Yufan Deng

Sanofi China

This presentation introduces the application of ICH E9(R1) estimand framework in vaccine clincal trials. Special considerations in vaccine efficay, immunogenicy and safety trials will be discussed in estimand framework to target the clinical questions of interest. The current implementation status of estimand framework in public disclosed vaccine trials, and the perception from regulatory agencies will also be reviewed.

A Bayesian phase I/II platform design with survival efficacy endpoint for dose optimization

Xian Shi, Jin Xu, ♦Rongji Mu

East China Normal University, East China Normal University, Shanghai Jiao Tong University

Motivated by a real-world drug development program, we propose a Bayesian phase I/II platform design to co-develop therapies with time-to-event efficacy endpoint (BPCT). We jointly model the toxicity outcome and the time-to-event efficacy outcome. At each interim, we update the dose-toxicity and dose-efficacy estimates, as well as the utility for risk-benefit tradeoffs, based on observed data from all indications. This approach informs indicationspecific decisions for dose escalation and de-escalation, and identifies the optimal biological dose for each indication. Simulation studies show that the proposed design has desirable operating characteristics, providing a highly flexible and efficient approach for dose optimization. The design has great potential to shorten the drug development timeline, save cost by reducing overlapping infrastructure, and expedite regulatory approval.

An overview of regional treatment effect evaluation via information borrowing in MRCTs

♦Yanghui Liu

Sanofi

Multi-regional clinical trials (MRCTs) have been widely used in contemporary drug development. Quantifying regional treatment effect is important for local registration in MRCTs. However, evaluating regional treatment effects based on regional data are usually inefficient due to limited sample sizes. This work provides an overview of methodological approaches to regional treatment effect evaluation through information borrowing across regions. For commonly used shrinkage estimation and Bayesian hierarchical models, simulation studies are conducted to illustrate their behaviors in different scenarios, in support for method and parameter selection in practice.

Timeline prediction in event driven clincial trials

♦Zhini Wang

In event-driven clinical trials, statistical power is primarily determined by the number of events. Consequently, analysis timing depends on event accrual throughout the study duration, making timeline prediction a practical issue. This presentation introduces statistical methodologies for this purpose and the comparison between different approaches.

25CHI097: Statistical analyses of several types of complex data

Copula-based models in compositional data analysis

Caikun Chen, Yu Fei,♦ Pengyi Liu, zhuo Chen

Department of Statistics, Yunnan University of Finance and Economics

Compositional data with unit–sum constraints often occurs in various fields. Neither the log-ratio method nor the Dirichlet regression offers a high level of interpretability and flexible fitting for such data since these models do not directly account for compositional data and often rely on assumptions of the same type of marginal distribution. The purpose of this paper is to address the limitations in interpretability and flexibility of existing regression models for compositional data. We propose a conditional distribution for compositional data as a response variable by incorporating the stick-breaking transformation and utilizing Beta, proportional inverse Gaussian (PIG) and simplex marginal distribution with the introduction of the Gaussian Copula function, which is more flexible and interpretable than the Multivariable Logit Normal and Dirichlet distributions. Besides, we utilize the Gaussian Newton-type algorithm for estimating the model parameters. The properties of the proposed estimator are assessed numerically through a simulation study and real-data analysis.

Growth curves mixture model for longitudinal data based on mean–covariance modeling

♦Yating Pan, Fangfang Pan, Jianxin Pan

Yunnan University of Finance and Economics, Yunnan University of Finance and Economics, Beijing Normal University, BNU-HKBU United International College

Growth curve mixture models account for unobserved heterogeneity by allowing the grouping matrix in the growth curve model to be predicted from the data. Existing methods usually model the mean in each subpopulation with assumption that observations sharing a common trajectory are independent or their covariance structure is pre-specified, but less research has explored modeling heterogeneous covariance structures. We introduce a joint model which models the mean and covariance structures simultaneously in a Gaussian mixture model within the framework of growth curve models, demonstrating how important the within-subject correlation is in clustering longitudinal data. Model parameters are estimated with an iteratively re-weighted least squares EM (IRLSEM) algorithm. We can identify different mean trajectories and covariance structures in all clusters. Simulations show that the proposed method performs well and gives more accurate clustering results by introducing covariance modeling. Real data analysis is also used to illustrate the usefulness of the proposed method.

Unified optimal model averaging with a general loss function based on cross-validation

♦Dalei Yu, Xinyu Zhang, Hua Liang

Xi'an Jiaotong University, University of Science and Technology of China and Academy of Mathematics and Systems Science, Chinese Academy of Sciences, George Washington University

Studying unified model averaging estimation for situations with complicated data structures,we propose a novel model averaging method based on cross-validation (MACV). MACV unifies a large class of new and existing model averaging estimators and covers a very general class of loss functions. Furthermore, to reduce the computational burden caused by the conventional leave-subject/one-out cross validation, we propose a SEcond-order-Approximated Leave-one/subject-out (SEAL) cross validation, which largely improves the computation efficiency. As a useful tool, we extend the Bernstein-type inequality for strongly mixing random variables that are not necessarily identically distributed. In the context of non-independent and non-identically distributed random variables, we establish the unified theory for analyzing the asymptotic behaviors of the proposed MACV and SEAL methods, where the number of candidate models is allowed to diverge with sample size. To demonstrate the breadth of the proposed methodology, we exemplify four optimal model averaging estimators under four important situations, i.e., longitudinal data with discrete responses, within-cluster correlation structure modeling, conditional prediction in spatial data, and quantile regression with a potential correlation structure. We conduct extensive simulation studies and analyze real-data examples to illustrate the advantages of the proposed methods.

25CHI099: Statistical Inference on high-dimensional covariance matrix

High-dimensional scale invariant discriminant analysis

♦Ming Li, Cheng Wang, Yanqing Yin, Shurong Zheng

Shandong Technology and Business University

In this paper, we propose a scale invariant linear discriminant analysis classifier for high-dimensional data with dense signals. The method is valid for both cases that the data dimension is smaller or greater than the sample size. Based on recent advances of the sample correlation matrix in random matrix theory, we derive the asymptotic limits of the error rate which characterizes the influences of the data dimension and the tuning parameter. The major advantage of our proposed classifier is scale invariant and it is applicable to any variances of the feature. Several numerical studies are investigated and our proposed classifier performs favorably in comparison to some existing methods.

Sparse estimation of high-dimensional cross-covariance matrices and its applications

♦Kazuyoshi Yata, Tetsuya Umino, Makoto Aoshima

University of Tsukuba, University of Tsukuba, University of Tsukuba

Sparse estimation of the entire covariance matrix has been studied extensively. However, research focused on estimating the cross-covariance matrix in high-dimensional settings is limited. In this talk, we propose a novel thresholding estimator of the cross-covariance matrix for high-dimension, low-sample-size (HDLSS) settings. We first investigate the asymptotic properties of the sample cross-covariance matrix and show that the estimator contains large amounts of noise in HDLSS setting, which renders it inconsistent. To overcome such difficulty, we develop a new thresholding estimator based on the automatic sparse estimation methodology and establish its consistency under mild conditions in HDLSS settings. Furthermore, we extend the proposed methodology to construct a thresholding estimator for high-dimensional mean vectors by applying the same principle used in the cross-covariance estimation. The performance of the proposed estimators is evaluated through empirical analysis using gene expression data.

Testing for large-dimensional covariance matrix under differential privacy

Shiwei Sang, ♦Yicheng Zeng, Shurong Zheng, Xuehu Zhu

Xi'an Jiaotong University, Sun Yat-sen University, Northeast Normal University, Xi'an Jiaotong University

The increasing prevalence of high-dimensional data across various applications has raised significant privacy concerns in statistical inference. In this paper, we propose a differentially private integrated test statistic for testing large-dimensional covariance structures, enabling accurate statistical insights while safeguarding privacy. First, we analyze the global sensitivity of sample eigenvalues for sub-Gaussian populations, where our method bypasses the commonly assumed boundedness of data covariates. For sufficiently large sample size, the privatized statistic guarantees privacy with high probability. Furthermore, when the ratio of dimension to sample size, $d/n \to y \in (0, \infty)$, the privatized test is asymptotically distribution-free with well-known critical values, and detects the local alternative hypotheses distinct from the null at the fastest rate of $1/\sqrt{n}$. Extensive numerical studies on synthetic and real data showcase the validity and powerfulness of our proposed method.

Approximate Normality in testing hierarchical covariance structures belonging to a quadratic subspace

Daniel Klein, ♦Yuli Liang, Mateusz John

P. J. Safarik University in Kosice, Slovakia, Guangxi Normal University, China, Institute of Mathematics, Poznan University of Technology, Poland

In this paper a hypothesis related the covariance structures belonging to a quadratic subspace under multivariate models is studied. The Rao score and the likelihood ratio test statistics are derived, and the exact distribution of the likelihood ratio test is determined. The results are applied to numerical illustrations.

25CHI101: Statistical Learning and Medical Diagnostics

Boundary Detection and Image Segmentation via Local Discrepancy Scan Statistics

♦Richeng Hu, Ngai Hang Chan, Chung Wang Wong, Chun Yip Yau

The Chinese University of Hong Kong, City University of Hong Kong, The University of Hong Kong, The Chinese University of Hong Kong

This paper proposes a novel three-step algorithm which effectively detects boundaries and segments noisy images by incorporating information from both local edges and global regions. Compared to existing methods, the proposed method does not require any initialization or predetermined number of regions, and allows intensity inhomogeneity within regions. In the algorithm, the first step scans out possible local change-boundaries on the image; the second step employs a novel nonparametric filament estimation to construct an over-segmented image; the final step develops a global criterion function to merge the over-segmented regions based on piecewise smooth models and obtain an optimal segmentation. Asymptotic properties including the uniform convergence rates for the proposed scan statistic and the consistency of the boundary estimation and the image segmentation are established. Extensive simulation experiments and a lung tumor data analysis are provided to illustrate the superb performance and wide applicability of the proposed algorithm.

AI-Powered Polyp Analysis in Colonoscopy: Improving Detection and Assessment

♦Jinfeng Xu

City University of Hong Kong

Precise detection, segmentation, and size estimation of colorectal polyps are vital for optimizing diagnostic accuracy and treatment strategies in colonoscopy. This study introduces an advanced AI-based framework to improve polyp analysis, integrating the Segment Anything Model (SAM) and fine-tuning as core elements within a broader methodological approach. The proposed system enhances polyp classification, segmentation, and size estimation from imaging data, providing critical insights for clinical decision-making. Evaluated on a dataset from Tuen Mun Hospital in Hong Kong's New Territories West Region, our method demonstrates superior performance compared to several established techniques, with significant gains in accuracy and reliability. These findings highlight the potential of this framework to advance AI-assisted colonoscopy, offering a scalable solution for improving patient outcomes through enhanced polyp assessment.

Predicting Future Change-points in Time Series

♦Chun Yip Yau

Chinese University of Hong Kong

Change-point detection and estimation procedures have been widely developed in the literature. However, commonly used approaches in change-point analysis primarily focus on detecting change-points within an entire time series (off-line methods), or the quickest detection of change-points in sequentially observed data (on-line methods). Both classes of methods are concerned with change-points that have already occurred. The arguably more important question of when future change-points may occur remains largely unexplored. In this paper, we develop a novel statistical model that describes the mechanism of change-point occurrence. Specifically, the model assumes a latent process in the form of a random walk driven by non-negative innovations, and an observed process which behaves differently when the latent process belongs to different regimes. By construction, an occurrence of a change-point is equivalent to crossing a regime threshold by the latent process. Therefore, by predicting when the latent process will cross the next regime threshold, future change-points can be forecasted. We establish probabilistic properties of the model such as stationarity and ergodicity, and develop a composite likelihood-based approach for parameter estimation and model selection. Moreover, we construct predictors and prediction intervals for future change-points based on the estimated model.

An application to EEG data will be illustrated.

Building an Artificial Intelligence-Based Infrastructure for Prospectively Validating Glaucoma Detection from 3D Optical Coherence Tomography Scans in real-world

♦*Anran Ran, Clement C. Tham, Carol Y. Cheung*

Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong

Purpose: Detecting glaucoma in the early stage is crucial for timely treatment and minimizing irreversible visual impairment. We have developed and retrospectively tested a 3D Artificial intelligence (AI) model to detect glaucomatous optic neuropathy (GON) from OCT scans. Here, we newly built an AI infrastructure to prospectively test the performance of GON detection from OCT scans using our 3D AI model in the real world.

Methods: We integrated our 3D AI model with an information management system and a commercially available OCT device as the AI infrastructure (Figure 1). This AI infrastructure included a user interface for real-time OCT image extraction, input data configuration, image uploading, image analysis via a graphics processing unit (GPU) server, and AI reports generation.

We recruited subjects at the Triage Unit of Hong Kong Eye Hospital prospectively. The inclusion criteria were: 1) subjects above 18 years old, and 2) referred by primary care settings due to glaucoma-related suspicious findings, such as increased cup-to-disc ratio, high intraocular pressure, and disc hemorrhage, or 3) referred by primary care settings for regular eye check-ups, or 4) with a family history of glaucoma, or 5) with diabetes mellitus. All the eligible subjects underwent OCT imaging, and the AI infrastructure analyzed their 3D OCT scans. Glaucoma specialists provided the reference standard for all the subjects.

We calculated the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and accuracy for performance evaluation.

Results: A total of 919 subjects were recruited and 167 of them were confirmed glaucoma. The AI infrastructure achieved an AUROC of 0.934, a sensitivity of 86.6%, a specificity of 88.3%, and an accuracy of 88.1% for GON assessment when compared with the reference standard. In addition, the AI infrastructure only took an average of 1-2 minutes to analyze the 3D OCT scan and generate an AI report for both eyes.

Conclusion: Our results demonstrated the AI infrastructure's good performance in a real-world clinical setting prospectively and provided evidence on silent mode evaluation. Further research, such as randomized control trials and health economic evaluation, is warranted before implementing the AI infrastructure in intended-use populations.

25CHI105: Statistical methods elevated by modern computation and massive data

Likelihood-based Nonparametric Receiver Operating Characteristic Curve Analysis in the Presence of Imperfect Reference Standard

♦*Peijun Sang, Yifan Sun, Qinglong Tian, Pengfei Li*

University of Waterloo, University of Waterloo, University of Waterloo, University of Waterloo

In diagnostic studies, researchers frequently encounter imperfect reference standards with some misclassified labels. Treating these as gold standards can bias receiver operating characteristic (ROC) curve analysis. To address this issue, we propose a novel likelihood-based method under a nonparametric density ratio model. This approach enables the reliable estimation of the ROC curve, area under the curve (AUC), partial AUC, and Youden's index with favorable statistical properties. To implement the method, we develop an efficient expectation-maximization algorithm algorithm. Extensive simulations evaluate its finite-sample performance, showing smaller mean squared errors in estimating the ROC curve, partial AUC, and Youden's index compared to existing methods. We apply the proposed approach to a malaria study.

Statistical Benefits when Incorporating LLM-Derived Predictions: Old Wine in a New Bottle?

♦*Jiwei Zhao*

University of Wisconsin - Madison

In biomedical studies involving electronic health records, manually extracting gold-standard phenotype

data is labor-intensive and limited in scale. The rise of generative AI, particularly large language models (LLMs), offers a systematic and significantly faster alternative through predictions, such as automated computational phenotypes (ACPs). However, directly substituting gold-standard data with these predictions, without addressing their differences, can introduce biases and lead to misleading conclusions. To address this challenge, we adopt a semi-supervised learning framework that integrates both labeled data (with gold-standard annotations) and unlabeled data (without gold-standard annotations) under the covariate shift paradigm. We propose doubly robust and semiparametrically efficient estimators to infer general target parameters. Through a rigorous efficiency analysis, we compare scenarios with and without the incorporation of LLM-derived predictions. Furthermore, we situate our approach within existing literature, drawing connections to prediction-powered inference and its extensions, as well as some seemingly unrelated concept such as surrogacy. To validate our theoretical findings, we conduct extensive synthetic experiments and apply our method to real-world data, demonstrating its practical advantages.

Robust Transfer Learning with Heterogeneous Data

Jing Wang, HaiYing Wang, ♦Kun Chen

University of Connecticut, University of Connecticut, University of Connecticut

Data fusion and transfer learning are rapidly growing fields that enhance model performance for a target population by leveraging other related data sources or tasks. The challenges lie in the various potential heterogeneities between the target and the external data, as well as various practical concerns that prevent a naïve data integration. We consider a realistic scenario where the target data is limited in size while the external data is large but contaminated with outliers; such data contamination, along with other computational and operational constraints, necessitate a proper selection or subsampling of the external data for transfer learning. We address this gap by studying robust transfer learning methods with subsamples of the external data, accounting for outliers deviating from the underlying true model. We provide non-asymptotic error bounds for several subsampling-based transfer learning estimators, clarifying the roles of sample sizes, signal strength, sampling rates, magnitude of outliers, and tail behaviors of error distributions, among other factors. Numerical studies show the superior performance of the proposed methods.

25CHI106: Statistical Methods for Survival Data with Complex Censoring and Missing or Mismeasured Covariates

Improving Estimation Efficiency for Case-cohort Studies with a Cure Fraction

♦Qingning Zhou, Xu Cao

University of North Carolina at Charlotte, University of California at Riverside

In the studies of time-to-event outcomes, it often happens that a fraction of subjects will never experience the event of interest, and these subjects are said to be cured. The studies with a cure fraction often yield a low event rate. To reduce cost and enhance study power, two-phase sampling designs are often adopted, especially when the covariates of interest are expensive to measure or obtain. In this paper, we consider the generalized case-cohort design for studies with a cure fraction. Under this design, the expensive covariates are measured for a subset of the study cohort, called subcohort, and for all or a subset of the remaining subjects outside the subcohort who have experienced the event during the study, called cases. We propose a two-step estimation procedure under a class of semiparametric transformation mixture cure models. We first develop a sieve maximum weighted likelihood method based only on the complete data and also devise an EM algorithm for implementation. We then update the resulting estimator via a working model between the outcome and cheap covariates or auxiliary variables using the full data. We show that the proposed update estimator is consistent and asymptotically at least as efficient as the complete-data estimator, regardless of whether the working model is correctly specified or not. We also propose a weighted bootstrap procedure for variance estimation. Extensive simulation studies demonstrate the superior performance of the proposed method in finite-sample. An application to the National Wilms' Tumor Study is provided for illustration.

Evaluating predictive accuracy of prognostic models with interval-censored data

♦Yang Qu, Yu Cheng

Central South University, China, University of Pittsburgh, USA

Evaluation methods of model predictive accuracy have been widely studied with right-censored outcomes. In practice, another type of censoring that is commonly encountered is interval censoring, where the occurrence of event of interest is only known to lie in some time intervals. Several methods studied evaluation problem in this setting, but often require modeling of the conditional survival function or the assessment process. In this project, we propose a new method to deal with evaluation of predictive accuracy of survival and competing risk models with interval-censored data, by using a simple imputation method. A simulation study is conducted to illustrate the finite-sample properties of the proposed method, and to investigate whether the proposed method can capture reduction in predictive accuracy resulting from exclusion of important variables. We further illustrate the performance of the proposed method in the real-world dataset.

A corrected smoothed score approach for semiparametric accelerated failure time model with error-contaminated covariates

♦Xiao Song

University of Georgia

We consider the semiparametric accelerated failure time (AFT) model with multiple covariates measured with error. Existing methods for the AFT model are either inconsistent,computationally intensive, or require stringent assumptions. To overcome these limitations, we develop a correction approach for a general smooth function of error-contaminated variables. We apply this method to the smoothed rank-based score function for the AFT model. The estimator is consistent and asymptotically normal. The finite-sample performance of the method is assessed by simulation studies. The approach is illustrated by application to data from an HIV clinical trial.

A Flexible Copula Model for Bivariate Survival Data with Dependent Censoring

Reuben Adatorwovor, ♦Yinghao Pan

University of Kentucky, University of North Carolina at Charlotte

Independent censoring is a key assumption usually made when analyzing time-to-event data. However, this assumption is untestable and can be problematic, particularly in studies with disproportionate loss to follow-up due to adverse events. This paper addresses the challenges associated with dependent censoring by introducing a likelihood-based approach for analyzing bivariate survival data under dependent censoring. A flexible Joe-Hu copula is used to formulate the interdependence within the quadruple times (two events and two censoring times). The marginal distribution of each event/censoring time is modeled via the Cox proportional hazards model. Our estimator possesses consistency and desirable asymptotic properties under regularity conditions. We provide results under extensive simulations with application to the Danish twin prostate cancer data.

25CHI108: Statistical Methods in Medical Applications

Heritability: a counterfactual perspective

Haochen Li, Jieru Shi, ♦Hongyuan Cao, Qingyuan Zhao

Florida State University, Cambridge University, Florida State University, Cambridge University

Nature vs nurture is a fundamental question in social, biological, and medical sciences. Motivated by twin studies, we propose a new notion that quantifies the relative contribution of nature and nurture from a causal inference perspective. We provide conditions for partial identification of this new notion with and without covariates in completely randomized experiments. In addition, we compare it with commonly used definitions of heritability in genetics through simulation studies. Our analysis sheds light on the important missing heritability problem.

Estimation and Prediction of Time-in-range (TIR) with Inpatient Continuous Glucose Monitoring

Qi Yu, Guillermo Umpierrez, ♦Limin Peng

Emory University, Emory University, Emory University

Continuous glucose monitoring (CGM) has been increasingly used in US hospitals for the care of patients with diabetes. Time-in-range (TIR), which measures the percent of time over a specified time window with glucose values within a target range, has served as a pivotal CGM-metric for assessing glycemic control. As Inpatient glycemia control generally involves clinical decisions (e.g., insulin adjustments) at sequential time points, dynamic prediction of TIR is of high interest to both clinicians and patients. However, this task is prone to multi-fold complications, which include a complex missing mechanism inherent to inpatient CGM data, the boundedness constraint to TIR which precludes straightforward regression modeling that typically requires some linearity assumption, and the presence of a large number of potential predictors. To address these challenges, we propose a random forest procedure which can accommodate nonlinear effects and interactions between predictors while simultaneously performing variable selection and dynamic prediction. Through utilizing a newly proposed nonparametric estimator of TIR, our proposal properly handles the complex data missingness associated with inpatient CGM data. Results from our numerical studies demonstrate the advantages of the proposed method over benchmark approaches.

Asymptotic distribution-free change-point detection for modern data based on a new ranking scheme

Doudou Zhou, ♦Hao Chen

National University of Singapore, University of California, Davis

Change-point detection (CPD) involves identifying distributional changes in a sequence of independent observations. Among nonparametric methods, rank-based methods are attractive due to their robustness and effectiveness and have been extensively studied for univariate data. However, they are not well explored for high-dimensional or non-Euclidean data. This paper proposes a new method, Rank INduced by Graph Change-Point Detection (RING-CPD), which utilizes graph-induced ranks to handle high-dimensional and non-Euclidean data. The new method is asymptotically distribution-free under the null hypothesis, and an analytic p-value approximation is provided for easy type-I error control. Simulation studies show that RING-CPD effectively detects change points across a wide range of alternatives and is also robust to heavy-tailed distribution and outliers. The new method is illustrated by the detection of seizures in a functional connectivity network dataset.

Investigating, Interpreting, and Optimizing Wearable Device Usage in Diabetes Patients

♦Jin Zhou, Bowen Zhang, Hua Zhou

University of California, Los Angeles , University of California, Los Angeles, University of California, Los Angeles

Continuous glucose monitoring (CGM) improves diabetes management, but adherence often wanes soon after initiation. We analyzed six-month wear data from 2,351 VA patients with T1D or T2D, modeling daily wear counts as empirical distributions and computing Wasserstein distances to capture trajectory dissimilarity. Spectral clustering identified three phenotypes—regular, fluctuating, and disengaged users. Using a double machine learning framework with four nuisance learners, we estimated each phenotype's causal impact on mean glucose and percent time-in-range. After one year, regular users saw a 2 mg/dL increase in mean glucose (95\% CI: 0.33–4.65; $p < 0.001$), whereas disengaged users experienced a 15.6 mg/dL increase (95\% CI: 7.2–23.9; $p < 0.001$). Sensitivity checks—varying clusters, nuisance methods, and propensity-score trimming—confirmed robustness. This coupling of optimal-transport clustering with DML uncovers meaningful adherence patterns and quantifies their graded effects on glycemic control.

25CHI112: Structured machine learning

Golden Ratio Weighting Prevents Model Collapse

Hengzhi He, ♦Shirong Xu, Guang Cheng

University of California, Los Angeles , University of California, Los Angeles, University of California, Los Angeles

Recent studies identified an intriguing phenomenon in recursive generative model training known as model collapse, where models trained on data generated by previous models exhibit severe performance degradation. Addressing this issue and developing more effective training strategies have become central challenges in generative model research. In this paper, we investigate this phenomenon theoretically within a novel framework, where generative models are iteratively trained on a combination of newly collected real data and synthetic data from the previous training step. To develop an optimal training strategy for integrating real and synthetic data, we evaluate the performance of a weighted training scheme in various scenarios, including Gaussian distribution estimation and linear regression. We theoretically characterize the impact of the mixing proportion and weighting scheme of synthetic data on the final model's performance. Our key finding is that, across different settings, the optimal weighting scheme under different proportions of synthetic data asymptotically follows a unified expression, revealing a fundamental trade-off between leveraging synthetic data and generative model performance. Notably, in some cases, the optimal weight assigned to real data corresponds to the reciprocal of the golden ratio. Finally, we validate our theoretical results on extensive simulated datasets and a real tabular dataset.

Two-way latent matching model for network analysis

♦*Ting Li, Jiangzhou Wang, Jianhua Guo*

The Hong Kong Polytechnic University, Shenzhen University, Beijing Gongshang University

We propose a Two-way Latent Matching Model (TLMM) for network data, designed to capture the intrinsic matching structures within networks. TLMM leverages dimension-reduction techniques from latent factor models and incorporates an additive structure to account for network symmetry. We establish conditions for model identifiability and develop the maximum likelihood method for parameter estimation, proving that the resulting estimators achieve optimal convergence rates and exhibit asymptotic normality as the network size grows. The effectiveness of the proposed algorithms and the asymptotic properties of the estimators are validated through extensive numerical simulations. Furthermore, applications to co-author network data and Divvy bicycle-sharing data demonstrate the model's practical effectiveness, yielding valuable insights.

Statistical Inference in Tensor Completion: Optimal Uncertainty Quantification and Statistical-to-Computational Gaps

Wanteng Ma, ♦*Dong Xia*

University of Pennsylvania, Hong Kong University of Science and Technology

We present a simple yet efficient method for statistical inference of tensor linear forms using incomplete and noisy observations. Under the Tucker low-rank tensor model and the missing-at-random assumption, we utilize an appropriate initial estimate along with a debiasing technique followed by a one-step power iteration to construct an asymptotically normal test statistic. This method is suitable for various statistical inference tasks, including constructing confidence intervals, inference under heteroskedastic and sub-exponential noise, and simultaneous testing. We demonstrate that the estimator achieves the Cramér-Rao lower bound on Riemannian manifolds, indicating its optimality in uncertainty quantification. We comprehensively examine the statistical-to-computational gaps and investigate the impact of initialization on the minimal conditions regarding sample size and signal-to-noise ratio required for accurate inference. Our findings show that with independent initialization, statistically optimal sample sizes and signal-to-noise ratios are sufficient for accurate inference. Conversely, if only dependent initialization is available, computationally optimal sample sizes and signal-to-noise ratio conditions still guarantee asymptotic normality without the need for data-splitting. We present the phase transition between computational and statistical limits. Numerical simulation results align with the theoretical findings.

Decentralized learning of low-rank matrix

Zihao Song, Weihua Zhao, ♦*Heng Lian*

Nantong University, Nantong University, henglian@cityu.edu.hk

Matrix factorization is a computationally efficient nonconvex approach for low-rank matrix recovery, utilizing an alternating minimization or a gradient descent algorithm, and its theoretical properties have been investigated in recent years. However, the behavior of the factorization-based matrix recovery problem in the decentralized setting is still unknown when data are distributed on multiple nodes. We consider the distributed algorithm and establish its (local) linear convergence up to the approximation error. Numerical results are also presented to illustrate the convergence of the algorithm over a general network.

25CHI114: Transforming clinical trials with causal inference thinking and methodology

A Connection Between Covariate Adjustment and Stratified Randomization in Randomized Clinical Trials

♦*Zhiwei Zhang*

Gilead Sciences

The statistical efficiency of randomized clinical trials can be improved by incorporating information from baseline covariates (i.e., pre-treatment patient characteristics). This can be done in the design stage using stratified (permutated block) randomization or in the analysis stage through covariate adjustment. This article makes a connection between covariate adjustment and stratified randomization in a general framework where all regular, asymptotically linear estimators are identified as augmented estimators. From a geometric perspective, covariate adjustment can be viewed as an attempt to approximate the optimal augmentation function, and stratified randomization improves a given approximation by moving it closer to the optimal augmentation function. The efficiency benefit of stratified randomization is asymptotically equivalent to attaching an optimal augmentation term based on the stratification factor. In designing a trial with stratified randomization, it is not essential to include all important c

An adaptive design for optimizing treatment assignment in randomized clinical trials

♦*Wei Zhang, Zhiwei Zhang, Aiyi Liu*

Chinese Academy of Sciences, Gilead Sciences, National Institutes of Health

The treatment assignment mechanism in a randomized clinical trial can be optimized for statistical efficiency within a specified class of randomization mechanisms. Optimal designs of this type have been characterized in terms of the variances of potential outcomes conditional on baseline covariates. Approximating these optimal designs requires information about the conditional variance functions, which is often unavailable or unreliable at the design stage. As a practical solution to this dilemma, we propose a multi-stage adaptive design that allows the treatment assignment mechanism to be modified at interim analyses based on accruing information about the conditional variance functions. This adaptation has profound implications on the distribution of trial data, which need to be accounted for in treatment effect estimation. We consider a class of treatment effect estimators that are consistent and asymptotically normal, identify the most efficient estimator within this class, and approximate the most efficient estimator by substituting estimates of unknown quantities. Simulation results indicate that, when there is little or no prior information available, the proposed design can bring substantial efficiency gains over conventional one-stage designs based on the same prior information. The methodology is illustrated with real data from a completed trial in stroke.

Incorporating external data for analyzing randomized clinical trials: A transfer learning approach

Yujia Gu, Hanzhong Liu, ♦Wei Ma

Renmin University of China, Tsinghua University, Renmin University of China

Randomized clinical trials are the gold standard for analyzing treatment effects, but high costs and ethical concerns can limit recruitment, potentially leading to invalid inferences. Incorporating external trial data with similar characteristics into the analysis using transfer learning appears promising for addressing these issues. In this paper, we present a formal framework for applying transfer learning to the analysis of clinical trials, considering three key perspectives: transfer algorithm, theoretical foundation, and inference method. For the algorithm, we adopt a parameter-based transfer learning approach to enhance the lasso-adjusted stratum-specific estimator developed for estimating treatment effects. A key component in constructing the transfer learning estimator is deriving the regression coefficient estimates within each stratum, accounting for the bias between source and target data. To provide a theoretical foundation, we derive the l1 convergence rate for the estimated regression coefficients and establish the asymptotic normality of the transfer learning estimator. Our results show that when external trial data resembles current trial data, the sample size requirements can be reduced compared to using only the current trial data. Finally, we propose a consistent nonparametric variance estimator to facilitate inference. Numerical studies demonstrate the effectiveness and robustness of our proposed estimator across various scenarios.

Joint Modeling of Longitudinal Biomarker and Survival Outcomes with the Presence of Competing Risk in Nested Case-Control Studies with Application to the TEDDY Microbiome Dataset

♦Jiyuan Hu

NYU Grossman School of Medicine

Background: Large-scale prospective cohort studies typically collect longitudinal biological samples alongside time-to-event outcomes to examine the association between biomarker dynamics and disease development. The nested case-control (NCC) design provides a cost-effective alternative to full cohort biomarker studies, particularly for rare and low-incidence diseases, while preserving statistical efficiency. Despite advances in joint modeling for longitudinal and time-to-event outcomes, few approaches address the unique challenges posed by NCC sampling, non-normally distributed biomarkers, and the presence of competing risk for the time-to-event outcomes.

Methods: Motivated by The Environmental Determinants of Diabetes in the Young (TEDDY) study, which utilizes an NCC design for microbial biomarker discovery, we propose a novel joint modeling approach, “JM-NCC”, tailored for NCC design with competing events. This framework integrates 1) a generalized linear mixed-effects model (longitudinal sub-model) to characterize biomarker trajectories over time, and 2) a cause-specific hazard model (survival sub-model) to link biomarker trajectories to competing events. Two maximum likelihood estimation (MLE) approaches, fJM-NCC and wJM-NCC are developed. The fJM-NCC approach fully utilizes longitudinal biomarker data from the NCC sub-cohort and survival and clinical metadata from the full cohort to construct the likelihood function. In contrast, the wJM-NCC approach uses only NCC sub-cohort data, incorporating inverse probability weighting to account for the NCC sampling process, and provides a flexible solution when full cohort metadata is unavailable.

Results: Simulation studies and application to the TEDDY microbiome dataset demonstrate the robustness and efficiency of the proposed methods. Both fJM-NCC and wJM-NCC yield unbiased parameter estimates and maintain well-controlled Type-I error rates across various scenarios. The fJM-NCC approach achieves statistical power comparable to the Oracle method, which models the longitudinal biomarkers, clinical metadata and survival outcomes from the full cohort. wJM-NCC ranks second and outperforms existing approaches—standard JM and conditional logistic regression (CLR)—in parameter estimation and hypothesis testing.

Conclusions: The proposed fJM-NCC and wJM-NCC methods are reliable and efficient tools for studying association between longitudinal biomarkers and competing events under NCC study design. These methods are well-suited for biomarker discovery in large-scale prospective studies, offering robust parameter estimation, well-controlled Type-I error rates, and satisfactory statistical power.

25CHI001: Advanced Experimental Design and Subsampling Approaches for Complex Data Analysis

Multi-resolution subsampling for linear classification with massive data

Haolin Chen, Holger Dette, ♦Jun Yu

Beijing Insititute of Technology, Ruhr-Universitat Bochum, Fakultat fur Mathematik, Beijing Insititute of Technology

Subsampling is one of the popular methods to balance statistical efficiency and computational efficiency in the big data era. Most approaches {aim to select} informative or representative sample points to achieve good overall information of the full data. The present work takes the view that sampling techniques are recommended for the region we focus on, and summary measures are enough to collect the information for the rest according to a well-designed data partitioning. We propose a subsampling strategy that collects global information described by summary measures and local information obtained from selected subsample points. Thus, we call it multi-resolution subsampling. We show that the proposed method leads to a more efficient subsample-based estimator for general linear classification problems. Some asymptotic properties of the proposed method are established, and connections to existing subsampling procedures are explored. Finally, we illustrate the proposed subsampling strategy via simulated and real-world examples.

A distance metric-based space-filling subsampling method for nonparametric models

Huaimin Diao, Dianpeng Wang, ♦Xu He

Shandong Technology and Business University, Beijing Institute of Technology, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Taking subset samples from the original data set is an efficient and popular strategy to handle massive data that is too large to be directly modeled. To optimize inference and prediction accuracy, it is crucial to employ a subsampling scheme to collect observations intelligently.In this paper, we propose a space-filling subsampling method that uses distance metric-based strata to select subsamples from high-volume data sets. To minimize the maximal distance from pairs of samples that locate in the same stratum, Voronoi cells of thinnest covering lattices are usedto partition the input space. In addition, subsamples that are space-filling according to the response are collected from each stratum. With the help of an algorithm to quickly identify the cell an observation locates in, the computational cost of our subsampling method is proportional to the number of observations and irrelevant to the number of cells, which makes our method applicable to extremely large data sets. Results from simulated studies and real data analysis show that the new method is remarkably better than existing approaches when used in conjunction with Gaussian process models.

Stratum Order-of-Addition Designs

Liushan Zhou, Ze Liu, Min-Qian Liu, ♦Guanzhou Chen

Nankai University, Nankai University, Nankai University, Nankai University

Order-of-addition experiments are widely employed in many fields of science and industry to study how the order of components being added influences the response. Although several classes of attractive order-of-addition designs have been proposed in recent years, the performance of these designs often relies on prespecified models and their run sizes are typically large unless the number of components is very small. In this paper, we put forward a model-free approach, called stratum order-of-addition designs, for order-of-addition experiments. By achieving stratum orthogonality of various strengths, the proposed designs are not only economical in run size, but also robust to model uncertainty. Theoretical justifications for these designs are established under a broad class of models for order-of-addition experiments. In addition, several deterministic methods are developed for constructing such designs with flexible run sizes. Numerical studies demonstrate that the proposed designs are competitive under many model-based and model-free criteria, as compared with existing designs in the literature.

Enhancing Sensitivity Analysis of Building Energy Performance through Batch-Sequential Maximum One-Factor-At-A-Time Designs

Qiang Zhao, Chunwei Zheng, Fasheng Sun, ♦Qian Xiao

Northeast Normal University, Nankai University, Northeast Normal University, Shanghai Jiao Tong University

Buildings consume a large share of the world's energy. Demand for cooling is growing as climate change increases the risk of overheating. Finding key design variables is critical to creating energy efficient buildings and retrofits. The MOFAT design performs small-sample sensitivity analysis efficiently and works well in many cases. However, in severe cold climates, the 5R1C model requires larger MOFAT designs, where a batch-sequential approach helps with limited computational resources. Creating this approach is difficult due to the strict design structure of MOFAT. This paper presents two novel batch-sequential MOFAT designs with flexible run sizes: I-MOFAT and S-MOFAT. We construct them using a new transformation function. Tests show that they outperform one-shot MOFAT and other methods, especially for sensitivity analysis of building energy performance in severe cold climates of China.

25CHI002: Advanced Learning Methods for Complex Medical Data

Parametric Modal Regression with Contaminated Covariates

Yanfei He, Jianhong Shi, ♦Weixing Song

Kansas State University

Modal regression provides a robust venue to describe how the response values of high frequency change with the covariates. In this talk, we propose a parametric modal regression procedure based upon the Gamma distribution family, when covariates are contaminated with normal measurement error. Compared to existing methods, the proposed procedure has three notable merits. First, it can handle multiple covariates subject to normal measurement errors; second, it possesses a simpler bias corrected likelihood function in general and a tractable expression in some special cases, resulting in faster computation and more precise estimation if the data distribution is correctly specified; third, empirical evidence shows that the Gamma-distribution-based modal regression has certain robustness against misspecification of the distribution of the measurement error. Numerical studies and real data applications are conducted to evaluate the finite sample performance of the proposed method.

Unsupervised Domain Adaptation with Adaptive f-Divergence: Tighter Variational Representation and Generalization Bounds

♦Fode Zhang,Yifan Zhu,Zhe Cheng

Southwestern University of Finance and Economics ,Southwestern University of Finance and Economics ,Southwestern University of Finance and Economics

Domain adaptation tackles domain shift by minimizing differences in data distributions between source and target domains, enhancing model performance on the target domain using source domain knowledge. The divergence measure in these distributions is crucial for adaptation. Most domain adaptation research uses a fixed divergence measure, which lacks flexibility for different data sets. We are interested in the scenarios where the divergence measure is dynamically adjustable in response to data set variations. This paper proposes an unsupervised domain adaptation framework predicated on a tighter variational representation of the -divergence measure, which can be selected synchronously instead of predetermined. The tighter variational representation of -divergence is obtained, which accelerates the statistical estimation of the divergence measure. The framework is theoretically supported, i.e., the target risk is shown to be bounded by the source risk, the discrepancy between the source domain and the target domain, and the total risk induced by the ideal hypothesis. Additionally, upper bounds pertaining to Rademacher complexity, covering number, and VC-dimensions are also derived. The experimental results demonstrate the proposed framework outperforms existing domain adaptation methods.

Large-scale survival analysis with a cure fraction

Bo Han, ♦Xiaoguang Wang, Liuquan Sun

Yunnan University, Dalian University of Technology, Institute of Mathematics and Systems Science, Chinese Academy of Sciences

With the advent of massive survival data with a cure fraction, large-scale regression for analyzing the effects of risk factors on a general population has become an emerging challenge. This article proposes a new probability-weighted method for estimation and inference for semiparametric cure regression models. We develop a flexible formulation of the mixture cure model consisting of the model-free incidence and the latency assumed by the semiparametric proportional hazards model. The susceptible probability assesses the concordance between the observations and the latency. With the susceptible probability as weight, we propose a weighted estimating equation method in a small-scale setting. Robust nonparametric estimation of the weight permits stable implementation of the estimation of regression parameters. A recursive probability-weighted estimation method based on data blocks with smaller sizes is further proposed, which achieves computational and memory efficiency in a large-scale or online setting. Asymptotic properties of the proposed estimators are established. We conduct simulation studies and a real data application to demonstrate the empirical performance of the proposed method.

Survival Prediction in ALS Patients Using Deep Learning

♦Haiyan Su, George Li, Liuxia Wang

Montclair State University, Carnegie Melon University, Afinity

Amyotrophic lateral sclerosis is a progressive neurodegenerative disease that affects nerve cells in the brain and spinal cord, affecting approximately 31,000 people in the United States. The FDA has approved several drugs for ALS that may reduce the rate of decline, or help manage symptoms. Currently, there is no known treatment that can halt or reverse the progression of ALS. A more accurate survival prediction may help for future clinical trial design to understand the progression of ALS and further prolong survival. In this study, we investigated the possibility of using deep learning to better predict survival for ALS patients using data from the PRO-ACT database. After developing the deep neural network model, it was compared with two models in the literature: the traditional statistical model using Cox Proportional Hazard (CPH) with ElasticNet, and a reliable machine learning model, Gradient Boo sting Machine (GBM). The comparison showed that deep learning model is comparable to GBM, and both models are superior to the CPH, in terms of prediction accuracy.

25CHI006: Advanced Statistical Methods for Spatial Transcriptomics

Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics

♦Xiang Zhou

University of Michigan

An essential task in spatial transcriptomics is identifying spatially variable genes (SVGs). Here, we present Celina, a statistical method for systematically detecting cell type-specific SVGs (ct-SVGs)—a subset of SVGs exhibiting distinct spatial expression patterns within specific cell types. Celina utilizes a spatially varying coefficient model to accurately capture each gene's spatial expression pattern in relation to the distribution of cell types across tissue locations, ensuring effective type I error control and high power. Celina proves powerful compared to existing methods in single-cell resolution spatial transcriptomics and stands as the only effective solution for spot-resolution spatial transcriptomics. Applied to five real datasets, Celina uncovers ct-SVGs associated with tumor progression and patient survival in lung cancer, identifies metagenes with unique spatial patterns linked to cell proliferation and immune response in kidney cancer, and detects genes preferentially expressed near amyloid-β plaques in an Alzheimer's model.

Cross-technology and cross-resolution framework for spatial omics annotation with CAESAR

♦Jin Liu, Xiao Zhang, Wei Liu

The Chinese University of Hong Kong (Shenzhen), The Chinese University of Hong Kong (Shenzhen), Sichuan University

The biotechnology of spatial omics has advanced rapidly over the past few years, with enhancements in both throughput and resolution. However, existing annotation pipelines in spatial omics predominantly rely on clustering methods and lack the flexibility to integrate extensive annotated information from single-cell RNA sequencing (scRNA-seq) due to discrepancies in spatial resolutions, species, or modalities. Here we introduce the CAESAR suite, an open-source software package that provides image-based spatial co-embedding of locations and genomic features. It uniquely transfers labels from scRNA-seq reference data, enabling the annotation of spatial omics datasets across different technologies, resolutions, species, and modalities, based on the conserved relationship between signature genes and cells/locations at an appropriate level of granularity. Notably, CAESAR enriches for location-level pathways, allowing for the detection of gradual biological pathway activation within spatially defined domain types. We demonstrate the advantages of CAESAR through a comprehensive analysis of five spatial omics datasets encompassing diverse technologies, resolutions, and modalities.

A de novo, spatially-aware and robust detection of phenotype-associated genes and tissue domains from multi-sample, multi-condition spatial transcriptomics

Wenlin Li, Yan Lu, Maocheng Zhu, Zhongkun Qu, Jin Liu, ♦Xiaobo Sun

School of Data Science, The Chinese University of Hong Kong-Shenzhen, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Statistics and Mathematics, Zhongnan University of Economics and Law, School of Data Science, The Chinese University of Hong Kong-Shenzhen, Department of Human Genetics, Emory University

The detection of genes and tissue domains associated with biological phenotypes from multi-sample and multi-condition spatial transcriptomics (ST) data provides critical insights into spatially resolved genomic alterations and cellular dynamics associated with pathological states. However, robust \textit{de novo} cross-sample DE analysis remains challenging, particularly in the presence of substantial inter-sample heterogeneity. Here, we present STANDEE, a novel contrastive masked autoencoder framework that generates batch-free gene representations through dynamic anchor gene alignment and spatial context preservation across tissue sections. STANDEE incorporates both vector quantization and rank consistency constraints to address embedding variability and data sparsity, with optional integration of histological features to enhance spatial gene representations. The framework enables multiple downstream analyses, including DE gene detection, disease-associated cell identification, disease-prone cell prediction, and missing gene imputation. Comprehensive benchmarks across datasets with diverse resolutions, modalities, and biological conditions demonstrate STANDEE's consistent superior performance over existing methods in revealing disease mechanisms and tissue dynamics.

Scaling up spatial transcriptomics for large-sized tissues

♦*Mingyao Li*

University of Pennsylvania

Recent advances in spatial transcriptomics (ST) technologies have transformed our ability to profile gene expression while retaining the crucial spatial context within tissues. However, existing ST platforms suffer from high costs, low resolution, limited gene coverage, and small tissue capture areas, restricting their use for large-scale investigations.Here we present iSCALE, a novel method that predicts super-resolution gene expression and automaticallyannotates cellular-level tissue architecture for whole-slide tissues, significantly exceeding the capture areas of standard ST platforms. The accuracy and robustness of iSCALE's predictions were confirmed through thorough evaluations, involving immunohistochemistry staining and pathologists' manual annotation. When applied to multiple sclerosis human brain samples, iSCALE revealed lesion associated cellular characteristics not captured by conventional ST experiments. We demonstrate its utility in analyzing large-sized tissues with automatic high throughput and unbiased tissue annotation, inferring cell type composition, and pinpointing regions of interest for features not discernible through human visual assessment alone.

25CHI007: Advancements in High-Dimensional Statistical Methods and Applications

Adaptive stratified sampling design in two-phase studies for average causal effect estimation

Min Zeng, Qiyu Wang, Zijian Sui, ♦Hong Zhang, Jinfeng Xu

City University of Hong Kong, University of Science and Technology of China, University of Science and Technology of China, University of Science and Technology of China, City University of Hong Kong

Causal inference using observational data suffers from numerous confounding effects, with greatly distorted average causal effect (ACE) estimates if the counfounders are ignored. Information on some confounders, such as genetic biomarkers and medical imaging, is prohibitively expensive to obtain in practice. Two-phase studies are resource-efficient solutions to this problem. In such studies, outcome, treatment, and inexpensive confounders are measured for a large number of subjects in the first phase; costly confounder measurements are then collected for a limited number of subjects in the second phase. An efficient statistical design is essential in controlling the cost arising in the second phase. In this paper, we propose an adaptive stratified sampling design (AdaStrat), which minimizes the variance of the ACE estimator with a given second-phase sample size. AdaStrat begins with gathering costly confounder measures for randomly selected pilot data, which are used to develop a stratification strategy and determine the sampling probabilities of strata. The resulting stratification and sampling strategy is applied to all first-phase subjects to determine the second-phase subjects with costly confounders measures. We rigorously show that AdaStrat produces a more efficient ACE estimator compared with the existing sampling designs with strata being pre-fixed. Finite sample properties of AdaStrat are evaluated through simulation studies, demonstrating its superiority against the fixed stratified sampling design (FixStrat), with relative efficiencies ranging from 20% to 30% in our simulation situations. The desired finite sample properties for AdaStrat are further confirmed through the application of the UK Biobank data.

High-dimensional statistical methods for analyzing three-dimensional genomic data

♦*Dechao Tian*

Sun Yat-sen University

The three-dimensional organization of the genome plays a crucial role in gene regulation and cellular function. Comparative analyses of Hi-C data and single-cell Hi-C data are critical to uncovering the genome's structure-function relationship in health and diseases. However, the high-dimensional, ultra-sparse, strongly dependent nature of these data poses significant statistical challenges. This talk will present recent collaborative works leveraging random matrix theory, statistical modeling, and deep generative models. Key progress include: 1) enhancing ultra-sparse single-cell Hi-C data, 2) testing differences between pairs of high-dimensional Hi-C contact matrices, and 3) identifying differential sub-matrices. Application to real data will be demonstrated, along with resulting biological discoveries, to demonstrate the potential of these methods in advancing genomic research.

High-Dimensional Bias Propagation in Multi-Temporal Covariance Dynamics: A Random Matrix Theory Approach

♦*Zhenggang Wang*

Southeast University

This study examines how estimation bias in sample covariance matrices affects portfolio construction in high-dimensional market settings, and how these effects propagate across multiple time horizons. Using random matrix theory, we analyze the bias characteristics with particular attention to how distortions affect the multi-temporal dynamics. The theoretical framework developed provides insights into the limitations of traditional portfolio optimization methods in high-dimensional settings and suggests potential approaches for novel analyzing procedures. This work contributes to the understanding of hierarchical relationships in financial covariance structures and their implications for portfolio management.

On generalized transformation models

♦*Zhezhen Jin*

Columbia University

In data analysis, generalized transformation models is useful to identify potential relationship between covariates and response. In this talk, I will review generalized transformation models and present their estimation and inference procedures. Some numerical studies will also be presented.

25CHI008: Advances in Causal Discovery for Omics Data

Robust Multi-ancestry PWAS Utilizing Bayesian

Fine-mapping

Chengli Zhang, Chong Wu, ♦ Haoran Xue

City University of Hong Kong, The University of Texas MD Anderson, City University of Hong Kong

Proteome-wide association studies (PWAS) have emerged as a powerful tool for identifying proteins associated with complex diseases, which can serve as potential drug targets. The conventional PWAS approach employs a two-stage least squares (2SLS) regression, utilizing genetic variants as instrumental variables (IVs). However, the validity of this approach can be compromised by the widespread pleiotropy of genetic variants, which can lead to the identification of false positive causal proteins for diseases. Furthermore, the varying linkage disequilibrium (LD) patterns and effect sizes of genetic variants across ancestries limit the power of PWAS when analyzing different populations separately. To address these challenges, we propose a robust and powerful PWAS method that integrates proteomics and diseases data from multiple ancestries and employs a Bayesian fine-mapping approach to detect invalid IVs. We demonstrate the effectiveness of our proposed method by applying it to large-scale biobank dataset, identifying putative causal proteins for complex human diseases.

A novel multivariable Mendelian randomization framework to disentangle highly correlated exposures with application to metabolomics

♦Lap Sum Chan, Mykhaylo Malakhov, Wei Pan

University of Minnesota, University of Minnesota, University of Minnesota

Mendelian randomization (MR) utilizes genome-wide association study (GWAS) summary data to infer causal relationships between exposures and outcomes, offering a valuable tool for identifying disease risk factors. Multivariable MR (MVMR) estimates the direct effects of multiple exposures on an outcome. This study tackles the issue of highly correlated exposures commonly observed in metabolomic data, a situation where existing MVMR methods often face reduced statistical power due to multicollinearity. We propose a robust extension of the MVMR framework that leverages constrained maximum likelihood (cML) and employs a Bayesian approach for identifying independent clusters of exposure signals. Applying our method to the UK Biobank metabolomic data for the largest Alzheimer disease (AD) cohort through a two-sample MR approach, we identified two independent signal clusters for AD: glutamine and lipids, with posterior inclusion probabilities (PIPs) of 95.0% and 81.5%, respectively. Our findings corroborate the hypothesized roles of glutamate and lipids in AD, providing quantitative support for their potential involvement.

Leveraging Cross-population Fine-mapping to Strengthen cis-Mendelian Randomization in TWAS

♦Mingxuan CAI

City University of Hong Kong

By integrating GWASs and resources from expression quantitative trait loci (eQTL) mapping studies, cis-Mendelian randomization (cis-MR) seeks to determine the causal effect of gene expression on human complex traits. However, two key challenges have hampered the accurate identification of causal genes. First, eQTLs inherited at a low recombination rate often harbor multiple causal variants in linkage disequilibrium (LD), making it difficult to identify independent instrumental variables (IVs), and limiting the power of cis-MR. Second, eQTL variants can affect the outcome trait through pathways other than the target gene (i.e. horizontal pleiotropy), which violates the MR assumption and leads to false positive results. Here, we introduce a statistical method called XMR that leverages cross-population fine-mapping to identify causal SNPs of gene expression as IVs for MR analysis, which helps maximize the MR power. At the same time, we explicitly correct for the horizontal pleiotropy by using genome-wide information, effectively controlling the type-I error in cis-MR.

Mitigating Subgroup Bias in Federated Selection of Sepsis Care Bundles

♦Yanyan Zhao, Peili Liu

Shandong University, Shandong University

The Surviving Sepsis Campaign (SSC) guidelines were developed to guide clinicians in treating patients with sepsis. However, concerns persist regarding their efficacy across diverse patients due to a predominant focus on data from White patients. This discrepancy suggests potential ethnic disparities in guideline development and implementation, which may affect their effectiveness globally. Additionally, amid growing concerns over data privacy, traditional distributed algorithms on raw data has hindered its reproducibility and generalizability due to data sharing constraints and communication costs. In our study, we aim to identify the statistically significant and ethnic-specific SSC guidelines associated with 28-day mortality, while addressing potential bias due to ethnic disparities and preserving data privacy. We propose a fair federated learning algorithm to mitigate ethnic bias during the selection process. We empirically demonstrate our method using data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) and eICU Collaborative Research Database (eICU-CRD) on sepsis patients.

25CHI011: Advances in Spatial Statistics with Random Field Modeling: Methods, Metrics, and Applications

Estimation and model selection in general spatial dynamic panel data

Li Hou, Baisuo Jin, ♦Yuehua Wu

University of Science and Technology of China, University of Science and Technology of China, York University

Common methods for estimating parameters of spatial dynamic panel data models include two-stage least squares, quasi-maximum likelihood, and generalized moments. In this talk, we present a method that uses the eigenvalues and eigenvectors of a spatial weight matrix to directly construct consistent least squares estimators of parameters of general spatial dynamic panel data models for both undirected and directed networks. Our method is conceptually simple and effective, and easy to implement. Results show that our parameter estimators are consistent and asymptotically normally distributed under mild conditions. We demonstrate the performance of our method through simulation studies and real data examples.

Local Maxima of Discrete Gaussian Processes

♦Dan Cheng, John Ginos

Arizona State University, Arizona State University

In this work, we derive the expected number and height distribution of local maxima for Gaussian processes defined on discrete parameter sets. We further demonstrate that as the discrete sets become increasingly dense, the expected number and height distribution will converge to those of the corresponding continuous Gaussian processes, respectively. Since real-world datasets are typically discrete, our findings offer valuable insights for statistical applications, including signal detection and change point detection.

A Bayesian nonstationary model for spatial binary data based on tree partition processes

♦Bohai Zhang, Furong Li, Jianxin Pan

Beijing Normal-Hong Kong Baptist University, Ocean University of China, Beijing Normal-Hong Kong Baptist University

Spatial binary data are ubiquitous nowadays in many disciplines, such as presence and absence of species in ecology, cloud mask/sea-ice detection in remote sensing, incidences of diseases in epidemiology, to name a few. Here, we propose a spatial generalized linear mixed model for spatial binary data, where spatial dependence is introduced by a latent Gaussian process. The nonstationarity of the latent process is achieved by domain partitioning method based on random spanning trees. We will give the exact Bayesian inference procedure of model parameters through data augmentation and equip the proposed method with fast computation strategy for large datasets. The performance of the proposed method is then demonstrated through simulation studies and applications to oceanographic datasets.

Time-varying vector random fields on the arccos-quasi-quadratic metric space

Juan Du, ♦Chunsheng Ma

Kansas State University, Wichita State University

An arccos-quasi-quadratic metric is defined on a subset of $\mathbb{R}^{d+1}$ such as a sphere, a ball, an ellipsoidal surface, an ellipsoid, a simplex, a conic surface, or a hyperbolic surface, and the corresponding metric space incorporates several important cases in a unified framework that makes possible for us to study metric-dependent random fields on different metric spaces in a unified manner. Over the arccos-quasi quadratic metric space, we construct a class of time-varying vector random fields via either spherical harmonics or ultraspherical polynomials, and build up various parametric and semiparametric covariance matrix structures.

The extension problem is discussed as well.

25CHI014: Advances in Statistical Methods and Applications

Stochastic feature selection with annealing and its applications to streaming data

♦Lizhe Sun, Adrian Barbu

Shanxi University of Finance and Economics, Florida State University

Feature selection is an important topic in high-dimensional statistics and machine learning, for prediction and understanding the underlying phenomena. It has many applications in computer vision, natural language processing, bioinformatics, etc. However, most feature selection methods in the literature have been proposed for offline learning, and the existing online feature selection methods have theoretical and practical limitations in true support recovery. In this presentation, we propose two novel online feature selection methods by stochastic gradient descent with a hard threshold operator. The proposed methods can simultaneously select the relevant features and build linear regression or classification models based on the selected variables. The theoretical justification is provided for the consistency of the proposed methods. Numerical experiments on simulated and real datasets show that the proposed methods compare favorably with state-of-the-art online methods from the literature.

A comparison of two models for detecting inconsistency in network meta-analysis

♦Lu Qin, Shishun Zhao, Wenlai Guo, Tiejun Tong, Ke Yang

Center for Applied Statistical Research and College of Mathematics,Jilin University,Changchun,China, Center for Applied Statistical Research and College of Mathematics,Jilin University,Changchun,China, Department of Hand Surgery,the Second Hospital of Jilin University,Changchun,China, Department of Mathematics,Hong Kong Baptist University,Hong Kong,China, Department of Statistics and Data Science, Beijing University of Technology, Beijing, China

The application of network meta-analysis is becoming increasingly widespread, and for a successful implementation, it requires that the direct comparison

result and the indirect comparison result should be consistent. Because of this,a proper detection of inconsistency is often a key issue in network metaanalysis as whether the results can be reliably used as a clinical guidance. Among the existing methods for detecting inconsistency, two commonly used models are the design-by-treatment interaction model and the side-splitting models. While the original side-splitting model was initially estimated using a Bayesian approach, in this context, we employ the frequentist approach. In this paper, we review these two types of models comprehensively as well as explore their relationship by treating the data structure of network meta-analysis as missing data and parameterizing the potential complete data for each model. Through both analytical and numerical studies, we verify that the side-splitting models are specific instances of the design-by-treatment interaction model,incorporating additional assumptions or under certain data structure. Moreover,the design-by-treatment interaction model exhibits robust performance across different data structures on inconsistency detection compared to the side-splitting models. Finally, as a practical guidance for inconsistency detection,we recommend utilizing the design-by-treatment interaction model when there is a lack of information about the potential location of inconsistency. By contrast, the side-splitting models can serve as a supplementary method especially when the number of studies in each design is small, enabling a comprehensive assessment of inconsistency from both global and local perspectives.

Distance-based Clustering of Functional Data with Derivative Principal Component Analysis

♦Ping Yu, Gongming Shi, Chunjie Wang, Xinyuan Song

Shanxi Normal University, Capital University of Economics and Business, Changchun University of Technology, The Chinese University of Hong Kong

Functional data analysis (FDA) is an important modern paradigm for handling infnite-dimensional data. An important task in FDA is clustering, which identifes subgroups based on the shapes of measured curves.Considering that derivatives can provide additional useful information about the shapes of functionals,we propose a novel L2 distance between two random functions by incorporating the functions and theirderivative information to determine the dissimilarity of curves under a unifed scheme for dense observations. The Karhunen–Loève expansion is used to approximate the curves and their derivatives. Cluster membership prediction for each curve intends to minimize the new distances between the observed and predicted curves through subspace projection among all possible clusters. We provide consistent estimators for the curves, curve derivatives, and the proposed distance. Identifability issues of the clustering procedure are also discussed. The utility of the proposed method is illustrated via simulation studies and applications to two real datasets. The proposed method can considerably improve cluster performance compared with existing functional clustering methods.

Testing for the equality of distributions in high dimension

♦Xu Li, Gongming Shi, Baoxue Zhang

Shanxi Normal University, Capital University of Economics and Business, Capital University of Economics and Business

In this paper, we propose a new homogeneous test for two high-dimensional random vectors. Our test is built on a new measure, the so-called characteristic distance, which can completely characterize the homogeneity of two distributions. The newly proposed metric has some desirable properties, for example, it possesses a clear and intuitive probabilistic interpretation, and can be used to address the high-dimensional distance inference. Theoretically, the limiting behaviors under the conventional fixed dimension and high-dimensional distance inference are thoroughly investigated. Simulation studies and real data analysis are presented to illustrate the nite-sample performance of the proposed test statistic.

25CHI017: Advances in Statistical Modeling: Variable Selection, Dependence, and Nonparametric Methods

BELIEF in Dependence

Benjamin Brown, ♦Kai Zhang, Xiao-Li Meng

UNC Chapel Hill, UNC Chapel Hill, Harvard University

Two linearly uncorrelated binary variables must be also independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, we develop a framework called binary expansion linear effect (BELIEF) for understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking Gaussian parallels. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how the choice of link affects modeling. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption as understood under their linear model counterparts. Furthermore, for a binary response, maximum likelihood estimation for GLMs paradoxically fails under complete separation, when the data are most discriminative, whereas BELIEF estimation automatically reveals the perfect predictor in the data that is responsible for complete separation. We explore these phenomena and provide related theoretical results. We also provide preliminary empirical demonstration of some theoretical results.

Variable selection for partially linear models and partially global Fréchet regression

♦Yichao Wu

University of Illinois Chicago

The first part of the talk will focus on the general partially linear model without any structure assumption on the nonparametric component. For such a model with both linear and nonlinear predictors being multivariate, we propose a new variable selection method. Our new method is a unified approach in the sense that it can select both linear and nonlinear predictors simultaneously by solving a single optimization problem. We prove that the proposed method achieves consistency.

The second part of the talk will be based on an ongoing research project. In this project, we are extending the above variable selection method to partially global Fréchet regression (Tucker and Wu, 2025 Statistica Sinica).

A NEW APPROACH TO SELECT LINEAR AND NONPARAMETRIC PREDICTORS SIMULTANEOUSLY FOR GENERALIZED PARTIALLY LINEAR MODELS

Youhan Lu, ♦Juan Hu, Yichao Wu

University of Illinois Chicago, DePaul University, University of Illinois Chicago

We introduce a novel approach for variable selection in the generalized partially linear model (GPLM) by building upon the previous research conducted by Lu et al. (2023). Our proposed method expands on their work by incorporating a local scoring algorithm. This approach enables the selection of linear and nonparametric predictors simultaneously by solving a single optimization problem. To showcase the effectiveness of our method in practice, we provide simulation examples involving logistic regression and Poisson regression models. Additionally, we present a real-world data example and engage in a discussion surrounding its application.

Quantile estimation for nonparametric regression models with autoregressive and moving average errors

♦Qi Zheng, Yunwei Cui

University of Louisville, Townson University

Quantile regression has emerged as a powerful and robust alternative to classical least-squares regression, particularly in time series forecasting where error terms may exhibit heavy tails. In this work, we consider the estimation of a nonparametric quantile regression model under a random design setup, allowing for correlated covariates and random errors that follow an autoregressive moving average (ARMA) process. To address the challenges posed by serial dependence and covariate correlation, we propose a spline-based method that jointly estimates the conditional quantile function and the ARMA parameters, rather than estimating model components sequentially. We establish the asymptotic properties of the proposed estimator under mild regularity conditions. Through extensive simulation studies, we demonstrate that our method performs well in finite samples and provides strong empirical support for the theoretical results. To illustrate the practical relevance of our approach, we apply it to model and forecast weekly natural gas consumption data in the state of Maryland across various quantile levels. This work adds a novel and flexible tool to the methodological arsenal for nonparametric quantile regression with serially correlated data.

25CHI021: Causal inference and decision-making

Proximal Inference on Population Intervention Indirect Effect

♦*Yang Bai, Yifan Cui, Baoluo Sun*

National University of Singapore, Zhejiang University, National University of Singapore

The population intervention indirect effect (PIIE) is a novel mediation effect representing the indirect component of the population intervention effect. Unlike traditional mediation measures, such as the natural indirect effect, the PIIE holds particular relevance in observational studies involving unethical exposures, when hypothetical interventions that impose harmful exposures are inappropriate. Although prior research has identified PIIE under unmeasured confounders between exposure and outcome, it has not fully addressed the confounding that affects the mediator. This study extends the PIIE identification to settings where unmeasured confounders influence exposure-outcome, exposure-mediator, and mediator-outcome relationships. Specifically, we leverage observed covariates as proxy variables for unmeasured confounders, constructing three proximal identification frameworks. Additionally, we characterize the semiparametric efficiency bound and develop multiply robust and locally efficient estimators. To handle high-dimensional nuisance parameters, we propose a debiased machine learning approach that achieves $\sqrt{n}$-consistency and asymptotic normality to estimate the true PIIE values, even when the machine learning estimators for the nuisance functions do not converge at $\sqrt{n}$-rate. In simulations, our estimators demonstrate higher confidence interval coverage rates than conventional methods across various model misspecifications. In a real data application, our approaches reveal an indirect effect of alcohol consumption on depression risk mediated by depersonalization symptoms.

Learning Robust Treatment Rules for Censored Data

Yifan Cui, Junyi Liu, ♦*Tao Shen, Zhengling Qi, Xi Chen*

Zhejiang University, Tsinghua University, National University of Singapore, George Washington University, New York University

There is a fast-growing literature on estimating optimal treatment rules directly by maximizing the expected outcome. In biomedical studies and operations applications, censored survival outcome is frequently observed, in which case the restricted mean survival time and survival probability are of great interest. In this paper, we propose two robust criteria for learning optimal treatment rules with censored survival outcomes; the former one targets at an optimal treatment rule maximizing the restricted mean survival time, where the restriction is specified by a given quantile such as median; the latter one targets at an optimal treatment rule maximizing buffered survival probabilities, where the predetermined threshold is adjusted to account the restricted mean survival time. We provide theoretical justifications for the proposed optimal treatment rules and develop a sampling-based difference-of-convex algorithm for learning them. In simulation studies, our estimators show improved performance compared to existing methods. We also demonstrate the proposed method using AIDS clinical trial data.

Causal mediation analysis of data fusion with application to bridging risk and relative efficacy of vaccines

♦*Pan Zhao, Oliver Dukes, Bo Zhang*

University of Cambridge, Ghent University, Fred Hutchinson Cancer Center

Refined vaccine regimens with variant-matched inserts are routinely approved by the regulatory agencies based on historical phase 3 clinical trials and immunobridging studies. Historical phase 3 clinical trials often help establish immune biomarkers that can reliably predict the risk or vaccine efficacy (VE) against a clinical endpoint. Once one or more immune correlates have been established, an immunobridging study, rather than another VE trial, will be conducted to compare the immunogenicity of an updated vaccine against that of an approved vaccine. In this article, we develop efficient and robust statistical methods that estimate the relative vaccine efficacy (relVE) of an updated vaccine versus an approved vaccine against the currently circulating strain, using relevant patient-level historical trials and immunobridging data. We discuss in detail identification assumptions, propose efficient and multiply robust estimators, and evaluate the finite sample performance of our proposed estimators. We demonstrate meaningful efficiency gain, which would translate to a smaller sample size when designing an immunobridging study, using our proposed estimators. We applied our framework to estimating the relative VE of multiple bivalent mRNA-1273 vaccines against monovalent prototype mRNA-1273 vaccine using data from the COVID-19 Variant Immunologic Landscape (COVAIL) Trial.

25CHI027: Frontier Statistical Methods for Single-cell RNA Sequencing Data

scTEL: Protein Expression Prediction in Single-cell Analysis Using Transformer

♦*Chaojie Wang*

Jiangsu University

CITE-seq provides a powerful method for simultaneously measuring RNA and protein expression at the single-cell level. The integrated analysis of RNA and protein expression in identical cells is crucial for revealing cellular heterogeneity. However, the high experimental costs associated with CITE-seq limit its widespread application. In this paper, we propose scTEL, a deep learning framework based on Transformer encoder layers, to establish a mapping from sequenced RNA expression to unobserved protein expression in the same cells. This computation-based approach significantly reduces the experimental costs of protein expression sequencing. We are now able to predict protein expression using single-cellRNAsequencing (scRNA-seq) data, which is well-established and available at a lower cost. Moreover, our scTEL model offers a unified framework for integrating multiple CITE-seq datasets, addressing the challenge posed by the partial overlap of protein panels across different datasets. Empirical validation on public CITE-seq datasets demonstrates scTEL significantly outperforms existing methods.

Large-scale imputation of spliced and unspliced RNA counts for Cell Lineage analysis

♦*Shanjun Mao*

Hunan University

RNA velocity has emerged as a powerful tool for inferring cellular differentiation trajectories by leveraging the temporal dynamics of spliced and unspliced RNA. However, current approaches often face limitations in scalability, robustness to noise, and integration of multi-modal data. Here, we present a novel framework that combines spliced/unspliced RNA counts with protein expression data to model transcriptional dynamics through a system of ordinary differential equations (ODEs), enhanced by deep learning architectures. Our method addresses large-scale imputation challenges to accurately estimate RNA velocity at single-cell resolution. By integrating protein expression as an additional regulatory layer, we capture post-transcriptional influences on cell fate decisions, thereby refining lineage inference. The approach employs neural networks to parameterize ODEs, facilitating scalable optimization across heterogeneous cell populations. We validate the framework on benchmark datasets, demonstrating improved robustness in reconstructing lineage trees at the cluster level compared to RNA-only models. The derived trajectories align with known developmental pathways and reveal transitional states obscured by conventional methods. This work advances multi-omics integration for lineage analysis, offering a principled strategy to dissect complex differentiation processes with applications in developmental biology and disease modeling.

Temporal mapping and clonal differentiation modelling from time-series single-cell RNA-seq data

Yijun Liu, Mingze Gao, ♦*Yuanhua Huang*

University of Hong Kong, University of Hong Kong, University of Hong Kong

International efforts have yielded extensive single-cell time-series atlas datasets, like those on mouse embryogenesis, providing a reference for mapping disease models across biomedical research. However, effectively using such data for temporal analysis of individual datasets is challenging due to the intricate nature of cell states and the tight coupling between time stamps and experimental batches. Here, we first introduce TemporalVAE, a deep generative model in a dual-objective setting that infers the biological time of each cell from a compressed latent space. Its accuracy and interpretability will be demonstrated in multiple scenarios, including embryogenesis, human disease, and cross-species mapping. Later, we will briefly introduce a related work on modelling clonal cell differentiation dynamics from lineage barcoded time-series scRNA-seq data.

25CHI030: Innovations and Partnerships in Data-Rich Environments: Emerging Advances in Matrix and Tensor Modeling

Shape Mediation Analysis in Alzheimer's Disease Studies

Xingcai Zhou, Miyeon Yeon, ♦*Jiangyan Wang, Shengxian Ding, Kaizhou Lei, Yanyong Zhao, Rongjie Liu, Chao Huang*

Nanjing Audit University

As a crucial tool in neuroscience, mediation analysis has been developed and widely adopted to elucidate the role of intermediary variables derived from neuroimaging data. Typically, structural equation models (SEMs) are employed to investigate the influences of exposures on outcomes, with model coefficients being interpreted as causal effects. While existing SEMs have proven to be effective tools for mediation analysis involving various neuroimaging-related mediators, limited research has explored scenarios where these mediators are derived from the shape space. In addition, the linear relationship assumption adopted in existing SEMs may lead to substantial efficiency losses and decreased predictive accuracy in real-world applications. To address these challenges, we introduce a novel framework for shape mediation analysis, designed to explore the causal relationships between genetic exposures and clinical outcomes, whether mediated or unmediated by shape-related factors while accounting for potential confounding variables. Within our framework, we apply the square-root velocity function to extract elastic shape representations, which reside within the linear Hilbert space of square-integrable functions. Subsequently, we introduce a two-layer shape regression model to characterize the relationships among neurocognitive outcomes, elastic shape mediators, genetic exposures, and clinical confounders. Both estimation and inference procedures are established for unknown parameters along with the corresponding causal estimands. The asymptotic properties of estimated quantities are investigated as well. Both simulated studies and real-data analyses demonstrate the superior performance of our proposed method in terms of estimation accuracy and robustness when compared to existing approaches for estimating causal estimands.

Matrix-factor-augmented regression

♦*Xiong Cai, Xinbing Kong, Xinlei Wu, Peng Zhao*

Nanjing Audit University, Southeast University, Nanjing Audit University, Jiangsu Normal University

As matrix-variate observations are increasingly available, to incorporate the interplay between the multi-cross-sections, we introduce a matrix-factor-augmented regression model (M-FARM) that proposes to predict ahead of time with factors of matrix predictors augmented in the regression. We show that the estimation error in the factor matrices, estimated by the projection procedure in the first step, enters into the estimation error of the regression parameters and the prediction error of the response variable with an asymptotically negligible rate. The central limit theorems of the estimates of the regression parameters are established under some mild conditions. Forecasting intervals with a theoretical guarantee are given. Monte-Carlo simulations justify the theoretical results. We find empirically that the augmented matrix factors do help in forecasting macroeconomic variables relative to the benchmark matrix autoregressive model and vector-factor-augmented regression model (V-FARM).

Matrix-quantile factor prediction for generalized matrix-variate regression

♦*Yongxin Liu*

Nanjing Audit University

This paper proposes a latent matrix-factor regression to predict responses that may come from an exponential distribution with high dimensional matrix-variates. The latent predictor is a low-dimensional factor extracted from the matrix quantile factor model, which provides a comprehensive and robust relationship between the high-dimensional covariates and low-rank factor predictor. Our prediction modeling conducts dimension reduction that respects the geometry characteristic of intrinsic two-dimensional structure of the matrix covariate. A two-step algorithm is used to estimate the matrix quantile factor model and the generalized regression. We establish the convergence rate of the estimated matrix coefficient and prediction. Extensive simulation studies show that the prediction capability of the proposed method outperform existing penalized methods and latent matrix model that only extract mean factors. An empirical application illustrates the usefulness of LaGMQFR by the COVID-19 data.

A Model-Based Monitoring Framework for Tensor Count Data in Passenger Flow Surveillance

♦*Yifan Li*

Nanjing Audit University

Tensor count data are increasingly prevalent across numerous applications, and passenger flow data in urban rail transit systems serve as a typical example. Real-time monitoring of passenger flow tensors is essential for identifying irregular behaviors and preventing severe consequences. However, existing online monitoring methods often fail to accommodate the unique characteristics of count data or are designed specifically for vectorized data, rendering them unsuitable for general-order tensor count processes. In this paper, we introduce a novel monitoring method, specifically designed for scenarios where count data appear in tensor form. The proposed method is based on a new Tensor Poisson Log-Normal Model. To address the estimation difficulties arising from the multi-dimensional latent variables in the model, we develop an efficient variational Gaussian approximation approach for Phase I modeling. In Phase II surveillance, an online parameter estimation algorithm is formulated based on the Laplace approximation method for the real-time computation requirements in online monitoring. Subsequently, we design an exponentially weighted likelihood based monitoring statistic to identify anomalies in online monitoring. Finally, we validate the effectiveness and superiority of our method through comprehensive simulations and apply it to real-time passenger flow surveillance in the Hong Kong Mass Transit Railway.

25CHI041: Integrate Statistics into Deep Learning for Digital Image Processing and Analysis

Solving Unbalanced Optimal Transport on Point Cloud by Tangent Radial Basis Function Method

♦*Jiangong Pan*

Tsinghua University

In this report, we solve unbalanced optimal transport (UOT) problem on surfaces represented by point clouds. Based on alternating direction method of multipliers algorithm, the original UOT problem can be solved by an iteration consists of three steps. The key ingredient is to solve a Poisson equation on point cloud which is solved by tangent radial basis function (TRBF) method. The proposed TRBF method requires only the point cloud and normal vectors to discretize the Poisson equation which simplify the computation significantly. Numerical experiments conducted on point clouds with varying geometry and topology demonstrate the effectiveness of the proposed method.

Learnable Mixture Distribution Prior for Deep Learning based Image Processing

♦*Jun Liu*

Beijing Normal University

The mixed distribution model is a model based on statistical classification. In this report, we start from solving the likelihood problem related to mixed distribution using the dual method, to illustrate its connection with mechanisms such as softmax activation, attention, and Transformer in deep learning. By constructing mixed distribution priors, we guide the algorithm and neural network design in data-driven image reconstruction and segmentation methods.

Learnable Nonlocal Self-similarity of Deep Features for Image Denoising

♦*Junying Meng, Faqiang Wang, Jun Liu*

Shanxi University, Beijing Normal University, Beijing Normal University

In this talk, we propose a learnable nonlocal self-similarity deep feature network for image denoising. Our method is motivated by the fact that the high-dimensional deep features obey a mixture probability distribution based on the Parzen-Rosenblatt window method. Then a regularizer with learnable nonlocal weights is proposed by considering the dual representation of the log-probability prior of the deep features. The proposed method provides a statistical and variational interpretation for the nonlocal self-attention mechanism. By adopting non-overlapping window and region decomposition techniques, we can significantly reduce the computational complexity of nonlocal self-similarity. The solution to the proposed variational problem can be formulated as a learnable nonlocal self-similarity network (LNSNet) for image denoising. This work offers a novel approach for constructing network structures that consider self-similarity and non-locality. Compared with several closely related denoising methods, the experimental results show the effectiveness of the proposed method in image denoising.

25CHI044: Machie learning for data assimilation

High-dimensional Ensemble Kalman Filter with Localization, Inflation and Iterative Updates

♦*Hao-Xuan Sun, Shouxia Wang, Xiaogu Zheng, Song Xi Chen*

Peking University, Beijing, China, Shanghai University of Finance and Economics, Shanghai, China, International Global Change Institute, Hamilton, New Zealand, Tsinghua University, Beijing, China

Accurate estimation of forecast error covariance matrices is an essential step in data assimilation, which becomes a challenging task for high-dimensional data assimilation. The standard ensemble Kalman filter (EnKF) may diverge due to both the limited ensemble size and the model bias. We propose to replace the sample covariance in the EnKF with a statistically consistent high-dimensional tapering covariance matrix estimator to counter the estimation problem under high dimensions. A high-dimensional EnKF scheme combining the covariance localization with the inflation method and the iterative update structure is developed. The proposed assimilation scheme is tested on the Lorenz-96 model with spatially correlated observation systems. The results demonstrate that the proposed method could improve the assimilation performance under multiple settings.

Generative Assimilation Forecasting

♦*Baoxiang Pan*

Institute of Atmospheric Physics, Chinese Academy of Science

Machine learning models have shown great success in predicting weather up to two weeks ahead, outperforming process-based benchmarks. However, existing approaches mostly focus on the prediction task, and do not incorporate the necessary data assimilation. Moreover, these models suffer from error accumulation in long roll-outs, limiting their applicability to seasonal predictions or climate projections. Here, we introduce Generative Assimilation and Prediction (GAP), a unified deep generative framework for assimilation and prediction of both weather and climate. By learning to quantify the probabilistic distribution of atmospheric states under observational, predictive, and external forcing constraints, GAP excels in a broad range of weather-climate related tasks, including data assimilation, seamless prediction, and climate simulation. In particular, GAP is competitive with state-of-the-art ensemble assimilation, probabilistic weather forecast and seasonal prediction, yields stable millennial simulations, and reproduces climate variability from daily to decadal time scales

Nonlinear assimilation with score-based sequential Langevin sampling

♦*Cheng Yuan*

Huazhong Normal University

This paper presents a novel approach for nonlinear assimilation called score-based sequential Langevin sampling (SSLS) within a recursive Bayesian framework. SSLS decomposes the assimilation process into a sequence of prediction and update steps, utilizing dynamic models for prediction and observation data for updating via score-based Langevin Monte Carlo. An annealing strategy is incorporated to enhance convergence and facilitate multi-modal sampling. The convergence of SSLS in TV-distance is analyzed under certain conditions, providing insights into error behavior related to hyper-parameters. Numerical examples demonstrate its outstanding performance in high-dimensional and nonlinear scenarios, as well as in situations with sparse or partial measurements. Furthermore, SSLS effectively quantifies the uncertainty associated with the estimated states, highlighting its potential for error calibration.

25CHI046: Modeling and inference for distributions and high dimensional data

Penalized weighted generalized estimation equations for high-dimensional longitudinal data with informative cluster size

♦*Haofeng Wang*

Hong Kong Baptist university

High-dimensional longitudinal data have become increasingly prevalent in recent studies, and penalized generalized estimating equations (GEEs) are often used to model such data. However, the desirable properties of the GEE method can break down when the outcome of interest is associated with cluster size, a phenomenon known as informative cluster size. In this article, we address this issue by formulating the effect of informative cluster size and proposing a novel weighted GEE approach to mitigate its impact, while extending the penalized version for high-dimensional settings. We show that the penalized weighted GEE approach achieves consistency in both model selection and estimation. Theoretically, we establish that the proposed penalized weighted GEE estimator is asymptotically equivalent to the oracle estimator, assuming the true model is known. This result indicates that the penalized weighted GEE approach retains the excellent properties of the GEE method and is robust to informative cluster sizes, thereby extending its applicability to highly complex situations. Simulations and a real data application further demonstrate that the penalized weighted GEE outperforms the existing alternative methods.

Two-sample tests for equal distributions in separable metric spaces: a unified semimetric-based approach

♦*Jin-Ting Zhang, Meichen Qian, Tianming Zhu*

National University of Singapore, National University of Singapore, Nanyang Technological University

With the advancement of data collection techniques, researchers frequently encounter complex data objects within separable metric spaces across various domains. One common interest lies in determining whether two groups of complex data objects originate from the same population. This paper introduces and examines a fast and accurate unified semimetric-based approach designed to tackle this challenge. The approach exhibits broad applicability across a wide range of research areas, such as bioinformatics, audiology, environmentology, finance, and more. It effectively identifies differences between the distributions of two complex datasets, including both high-dimensional data and functional data. The asymptotic null and alternative distributions of the proposed test statistic are established. Unlike the permutation approach, a unified, rapid and precise method to approximate the null distribution is described. Furthermore, the proposed test is shown to be root-n consistent. Numerical results are presented for illustrating the excellent performance of the proposed test in terms of size control, power, and computational cost. Additionally, the applications of the proposed test are showcased through examples involving both high-dimensional data and functional data.

Generalized Median of Means Principle for Bayesian Inference

Stanislav Minsker, ♦*Shunan Yao*

University of Southern California, Hong Kong Baptist University

The topic of robustness is experiencing a resurgence of interest in the statistical and machine learning communities. In particular, robust algorithms making use of the so-called median of means estimator were shown to satisfy strong performance guarantees for many problems, including estimation of the mean, covariance structure as well as linear regression. In this work, we propose an extension of the median of means principle to the Bayesian framework, leading to the notion of the robust posterior distribution. In particular, we (a) quantify robustness of this posterior to outliers, (b) show that it satisfies a version of the Bernstein-von Mises theorem that connects Bayesian credible sets to the traditional confidence intervals, and (c) demonstrate that our approach performs well in applications.

Oracle-efficient estimation and trend inference in

non-stationary time series with trend and heteroscedastic ARMA error

♦*Chen Zhong*

Fuzhou University

The non-stationary time series often contain an unknown trend and unobserved error terms. The error terms in the proposed model consist of a smooth variance function and the latent stationary ARMA series, which allows heteroscedasticity at different time points. The theoretically justified two-step B-spline estimation method is proposed for the trend and variance function in the model, and then residuals are obtained by removing the trend and variance function estimators from the data. The maximum likelihood estimator (MLE) for the latent ARMA error coefficients based on the residuals is shown to be oracally efficient in the sense that it has the same asymptotic distribution as the infeasible MLE if the trend and variance function were known. In addition to the oracle efficiency, a kernel estimator is obtained for the trend function and shown to converge to the Gumbel distribution. It yields an asymptotically correct simultaneous confidence band (SCB) for the trend function, which can be used to test the specific form of trend. A simulation-based procedure is proposed to implement the SCB, and simulation and real data analysis illustrate the finite sample performance.

25CHI059: Novel Machine Learning Methods for Disease Progression and Precision Medicine

Dynamic System for Modeling Latent Disease Progression and Treatment Effect

♦*Zexi Cai, Yuanjia Wang, Shanghong Xie*

Columbia University, Columbia University, University of South Carolina

Neurodegenerative diseases, such as Alzheimer's disease, pose significant public health risks due to their chronic nature. The progressive decline in cognitive function is difficult to analyze due to the heterogeneity of the disease progression without a common reference timeline. In this paper, we propose a dynamic disease progression model leveraging ordinary differential equations to characterize the latent dynamics of neurodegenerative disorders and assess the impact of therapeutic interventions. By integrating temporal registration functions tailored to individual patient characteristics, our model accounts for individual variability in disease progression, and enables better alignment of patient-specific trajectories to a common time frame. Both parametric and nonparametric forms of ordinary differential equations are considered. Parameter estimation is achieved through a variational inference approach due to the unknown initial conditions, and the registration functions are associated with baseline characteristics and estimated using a neural network architecture. Simulation studies demonstrate the model's effectiveness in recovering the underlying parameters and predicting future progression. We further apply our method to the Alzheimer's Disease Neuroimaging Initiative study, where the proposed model shows adequate performance in modeling multiple markers of cognitive decline with longitudinal time registration, and reveals a higher risk of rapid deterioration in cognition for female patients and carriers of APOE mutation.

LATENT GAUSSIAN PROCESS JOINT MODEL FOR INTEGRATIVE ANALYSIS OF MULTIMODAL BIOMARKERS AND INITIATION OF MEDICATION OF PARKINSON'S DISEASE

Junxuan Chen, Xiangnan Feng, ♦*Kai Kang*

Sun Yat-sen University, Fudan University, Sun Yat-sen University

Patients affected by Parkinson's disease (PD) usually experience a range of symptoms, including movement disorder, sleep disturbance and brain structural changes, and may start treatment at certain time points throughout clinical course. Current research focusing on a single modality (e.g., movement disorder) fails to display the full picture of the disease progression and its connection with timing of symptomatic therapy. In this paper, we integrate longitudinal mixed types of measurements from multiple modalities (e.g., binary/ordinal clinical measures, continuous neuroimaging biomarkers) and uncover their relationship with the initiation of PD treatment using a latent Gaussian process joint model. The dependence structure between observed multimodal biomarkers is characterized by several underlying unobserved Gaussian processes through a generalized factor model. Instead of assuming a certain parametric form a priori, the Gaussian processes, with its philosophy to let data speak for themselves, are expected to capture the possibly nonlinear and complex progression profiles of PD-related measurements. The obtained Gaussian processes are also incorporated into a varying coefficient Cox model so as to jointly monitor the longitudinal PD-related measurements and their time-varying effects on time-to-initiation of PD treatment. The application of our proposed method to the Parkinson's Progression Markers Initiative dataset yields a comprehensive understanding of PD progression by synthesizing information from clinical, biological, and neuroimaging perspectives.

Heterogeneous Quantile Treatment Effect Estimation with High-Dimensional Confounding

♦*Huichen Zhu*

The Chinese University of Hong Kong

Understanding heterogeneous treatment responses is essential for advancing precision medicine. This is because individuals often respond differently to the same treatment due to their unique characteristics and circumstances, including patient demographics, genetic predispositions, and environmental exposures. Moreover, inferring causal relationships or associations from observational data can be compromised by the presence of confounding factors, which can sometimes be high-dimensional. In this paper, we focus on estimating heterogeneous quantile treatment effects in the context of high-dimensional confounding. Our innovative approach leverages quantile regression and random forests to capture the variability of treatment effects across both covariates and outcome distributions. Additionally, we employ orthogonal estimation equations to robustly adjust for high-dimensional confounding. We rigorously explore the theoretical properties of our proposed estimator and demonstrate its finite-sample performance through comprehensive simulations. By addressing these complexities, our work aims to enhance the reliability of treatment effect estimates, ultimately contributing to more personalized and effective medical interventions.

25CHI061: Observational data analysis with complex study designs

Stochastic Explicit Calibration Algorithm for Survival Models

Jeongho Park

Yonsei University

Calibration is essential for risk evaluation in various fields; including medicine, finance, and reliability analysis. Although extensive research has focused on calibration in classification and regression tasks using deep neural networks, survival analysis remains relatively underexplored, resulting in the lack of improved calibration methods. In particular, while previous work has proposed a calibration method for survival analysis, it relies on fixed bins, which can lead to biased calibration assessments and substantial loss of predictive accuracy in pursuit of calibration. This gap can hinder an accurate assessment of survival functions, leading to increased risk management costs. In this study, we introduce Stochastic Explicit Calibration (S-cal), an algorithm that employs random intervals instead of fixed bins, thereby advancing the calibration methods used in deep networks. The calibration performance of S-cal is evaluated using metrics specifically designed for handling censored data, such as D-calibration and the Kolmogorov-Smirnov metric. Extensive experiments on synthetic and real-world datasets demonstrate that S-cal consistently outperforms existing methods in terms of calibration accuracy. In addition, we highlight how improved calibration can improve downstream tasks, including optimizing resource allocation and improving patient care decisions. This work presents a significant advancement in the study of calibration for survival analysis, offering valuable information for more reliable risk assessment models.

Leveraging LLM-Derived Gene Embeddings for Gene-Expression Analysis

♦*Jun Li*

University of Notre Dame

Large language models (LLMs) have introduced new possibilities for analyzing gene expression data. One promising approach involves transforming textual descriptions of genes—such as those provided by NCBI—into dense numerical embeddings using LLMs. This presentation explores what functional information is captured by these embeddings, how they can enhance the analysis of single-cell gene expression data, and how they can contribute to predicting transcriptomic outcomes in gene perturbation experiments, such as those conducted with Perturb-seq.

Exponential Power Mixture of Experts Model: Estimation, Clustering, and Variable Selection

♦*Zhenghui Feng, Xuefei Qi, Heng Peng, Xingbai Xu, Jie Xue*

Harbin Institute of Technology, Xiamen University, HongKong Baptist University, Xiamen University, Xiamen University

Abtract: The mixture of experts model is a popular framework for modeling heterogeneity in data for regression, classification, and clustering. For regression and cluster analyses of continuous data, MoE usually uses normal experts following the Gaussian distribution. However, for a set of data containing a group or groups of observations with heavy tails or outliers, the use of normal experts is unsuitable and can unduly affect the fit of the MoE model. Motivated by the flexibility of the exponential power distribution, we propose EPMoE model and provide the corresponding estimation method, taking into account variable selection and the theoretical properties of the model. Numerical simulations demonstrate the effectiveness of our proposed method and algorithm, and finally, we provide a real data analysis to illustrate the application of our approach.

Kernel Ridge Regression with Predicted Feature Inputs and Applications to Factor-Based Nonparametric Regression

Xin Bing, Xin He, ♦*Chao Wang*

University of Toronto, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

Kernel methods, particularly kernel ridge regression (KRR), are time-proven, powerful nonparametric regression techniques known for their rich capacity, analytical simplicity, and com

putational tractability. The analysis of their predictive performance has received continuous attention for more than two decades. However, in many modern regression problems where the feature inputs used in KRR cannot be directly observed and must instead be inferred from other measurements, the theoretical foundations of KRR remain largely unexplored. In this paper, we introduce a novel approach for analyzing KRR with predicted feature inputs. Our framework is not only essential for handling predicted feature inputs—enabling us to derive risk bounds without imposing any assumptions on the error of the predicted features—but also strengthens existing analyses in the classical setting by allowing arbitrary model misspecification, requiring weaker conditions under the squared loss, particularly allowing both an unbounded response and an unbounded function class, and being flexible enough to accommodate other convex loss functions. We apply our general theory to factor-based nonparametric regression models and establish the minimax optimality of KRR when the feature inputs are predicted using principal component analysis. Our theoretical findings are further corroborated by simulation studies.

25CHI063: Random matrices and high-dimensional statistics

Eigenvalues of large dimensional information plus noise type matrices

♦*Huanchao Zhou, Zhidong Bai, Jiang Hu, Jack Silverstein*

School of Mathematics and Computational Science, Xiangtan University,, School of Mathematics and Statistics, Northeast Normal University, School of Mathematics and Statistics, Northeast Normal University, Department of Mathematics, North Carolina State University

Let B_n = \frac{1}{n} \left(R_n + T_n^{1/2} X_n \right) \left(R_n + T_n^{1/2} X_n \right)^*, where X_n is a p×n matrix with independent standardized random variables, R_n is a p×n non-random matrix and T_n is a p×p non-random, nonnegative definite Hermitian matrix. The matrix B_n is referred to as the information-plus-noise type matrix, where R_n contains the information and T_n^{1/2} X_n is the noise matrix with the covariance matrix T_n. This talk will introduce some recent asymptotic global results concerning the eigenvalues of the B_n under high dimensional regimes. Key topics will include the convergence behaviors of these eigenvalues and the conditions under which these behaviors hold. We will particularly focus on the limiting spectral distribution of B_n as both matrix dimensions p and n grow to infinity, and how this limit is

influenced by the interplay between the information matrix Rn and the noise covariance matrix T_n.

On spiked eigenvalues of a renormalized sample covariance matrix from multi-population

Weiming Li, Zeng Li, ♦Junpeng Zhu

Shanghai University of Finance and Economics, Southern University of Science and Technology, Southern University of Science and Technology

Sample covariance matrices from multi-population typically exhibit several large spiked eigenvalues, which stem from differences between population means and are crucial for inference on the underlying data structure. This paper investigates the asymptotic properties of spiked eigenvalues of a renormalized sample covariance matrices from multi-population in the ultrahigh dimensional context where the dimension-to-sample size ratio $p / n \rightarrow \infty$. The first- and second-order convergence of these spikes are established based on asymptotic properties of three types of sesquilinear forms from multipopulation. These findings are further applied to two scenarios, including determination of total number of subgroups and a new criterion for evaluating clustering results in the absence of true labels. Additionally, we provide a unified framework with $p / n \rightarrow c \in(0, \infty]$ that integrates the asymptotic results in both high and ultrahigh dimensional settings.

Limiting spectral distribution for a cross data matrix-based matrix

♦Shao-Hsuan Wang

National Central University

The concept of the cross-data matrix originates from the work of Yata and Aoshima (2010),who demonstrated that the cross-data matrix-based principal component analysis (CDM-PCA) method can effectively reduce noise and enhance the performance of principal component anal- ysis (PCA) in high-dimensional, low-sample-size settings. This innovative approach has in-spired numerous subsequent studies. For instance, Wang, Huang, and Chen (2020) established the asymptotic normality of estimates for principal component directions, while Wang and Huang (2022) derived finite-sample approximations and explored the asymptotic behavior of CDM-based PCA through matrix perturbation theory. More recently, Hung and Huang (2023) introduced a more stable variant of CDM-PCA, termed product-PCA (PPCA). This formulation offers a more convenient structure for theoretical analysis and has been shown to be more robust than PCA in preserving the correct ordering of leading eigenvalues, even in the presence of outliers.

In this talk, I will discuss recent advances in the cross-data matrix-based methods for

high-dimensional data analysis. Moreover, I will present the limiting spectral distribution (LSD)

for the singular values of large cross-data matrix-based sample covariance matrix. Additionally,

I will compare this distribution with the Marchenko–Pastur law (MP law), which characterizes

the asymptotic behavior of the singular values of a large sample covariance matrix.

25CHI069: Recent Advances in Microbiome Data Analysis

Knockoff-based high-dimensional mediator identification and its application in microbiome research

♦Tiantian Liu, Dong Xu

China Pharmaceutical University, Shanghai Jiaotong University

Motivation: Microbiome data is characterized by high dimensionality, sparsity, and compositionality, posing significant challenges for studying microbial communities and their interactions with hosts and the environment. Particularly in high-dimensional settings, achieving variable selection for causal mediation analysis has become a critical challenge.

Results: This study proposes a novel microbial mediation analysis method, KAMA, to identify and assess the mediating role of microbes in the relationship between environmental exposures or interventions and host outcomes. By integrating the KnockoffScreen generator with an aggregation of multiple knockoff (AKO) approach, KAMA significantly enhances the detection capability of mediation effects in high-dimensional microbiome data. Comprehensive simulations demonstrate that KAMA exhibits superior performance in FDR control, statistical power, and result stability compared to existing mediation analysis methods. Furthermore, two real-data applications illustrate the effectiveness and practicality of KAMA. Our findings suggest that KAMA is a powerful tool for elucidating the mediating role of the microbiome and deepening our understanding of the complex factors that influence host health.

gmmcoda: Graphical model for the mixture of compositional data and absolute abundance data with applications to microbiome studies

Shen Zhang, ♦Huaying Fang, Tao Hu

Capital Normal University, Capital Normal University, Capital Normal University

Probabilistic graphical models provide efficient approaches to exploring the relationship of variables in various applications. Gaussian graphical model (GGM) is popular for constructing the conditional dependence network of interested variables. However, GMM is inappropriate in some applications such as microbiome studies, in which only relative abundances (referred to as compositional data in statistics) can be observed for some variables. Recently, some algorithms have been proposed to deal with the graphical modeling problem for compositional data. Nevertheless, there is a lack of statistical methods for inferring the interaction network for the mixture of compositional data and absolute abundance data. In this study, we propose a probabilistic graphical model for modeling interactions of variables for the mixture of compositional data and absolute abundance data. A novel maximum penalized likelihood estimator, called gmmcoda, is introduced for inferring the network from the mixture data. We develop a majorization-minimization algorithm to solve the optimization problem involved in gmmcoda. The performance of gmmcoda is evaluated and compared with other existing methods by simulation studies. Additionally, we apply gmmcoda to one microbiome data including microbial abundance as compositional data and gene expression data as absolute abundance data. The microbe-gene interactions detected by gmmcoda are further validated using previous studies.

Integrative analysis of microbial 16S gene and shotgun metagenomic sequencing data improves statistical efficiency

in testing differential abundance

♦Yicong Mao, Ye Yue, Timothy Read, Veronika Fedirko, Glen Satten, Xuan Chen, Xiang Zhan, Yi-Juan Hu

Department of Biostatistics, Peking University, Department of Biostatistics and Bioinformatics, Emory University, Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Department of Epidemiology, University of Texas MD Anderson Cancer Center & Department of Epidemiology, Emory University, Department of Gynecology and Obstetrics, Emory University School of Medicine, College of Economics and Management, Huazhong Agricultural University, School of Statistics and Data Science, Southeast University, Department of Biostatistics, School of Public Health, Peking University & Beijing International Center for Mathematical Research, Peking University & Center for Statistical Science, Peking University

The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing and shotgun metagenomic sequencing. Surprisingly, many microbiome studies have performed both experiments on the same cohort of samples. The two sequencing datasets often reveal consistent patterns of microbial signatures, suggesting that an integrative analysis of both datasets could enhance the testing power for these signatures. However, differential experimental biases, partially overlapping samples, and uneven library sizes pose tremendous challenges when combining the two datasets. In this article, we introduce the first method of this kind, named Com-2seq, that combines the two datasets for testing differential abundance at the genus level as well as the community level while overcoming these difficulties. Our simulation studies demonstrate that Com-2seq substantially enhances statistical efficiency over analysis of a single dataset and outperforms two ad hoc approaches to integrative analysis. In analysis of real microbiome data, Com-2seq uncovered scientifically plausible findings, namely, the association of Butyrivibrio , Gemella and Ignavigranum with prediabetes status, which would have been missed by analyzing a single dataset. Butyrivibrio failed to reach the significance level in the analysis of each dataset despite showing a consistent trend; Gemella and Ignavigranum failed to produce adequate data in the 16S experiment.

Differential abundance analysis of sequence count data

Guanxun Li, Xianyang Zhang, ♦Huijuan Zhou

Beijing Normal University at Zhuhai, Texas A&M Univerisity, Shanghai University of Finance and Economics

The development of high-throughput sequencing technologies has advanced microbiome and single-cell studies, with the resulting data typically summarized as a feature-by-sample count matrix. The inherent compositionality of sequence count data presents a major challenge for differential abundance analysis. Some recent studies have proposed incorporating sequencing depth as an offset term in their regression frameworks; however, this approach does not fully address the compositional nature of the data. In our study, we explicitly account for compositionality within the model, resulting in a more concise and structured framework that is well-suited for both microbiome and single-cell count data.

25CHI073: Recent Advances in Single-cell Data Analysis

GraphPCA: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data

Jiyuan Yang, Lu Wang, Lin Liu, ♦Xiaoqi Zheng

Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University School of Medicine,, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai Jiao Tong University School of Medicine

The rapid advancement of spatial transcriptomics technologies has revolutionized our understanding of cell heterogeneity and intricate spatial structures within tissues and organs. However, the high dimensionality and noise in spatial transcriptomic data present significant challenges for downstream data analyses. Here, we develop GraphPCA, an interpretable and quasi-linear dimension reduction algorithm that leverages the strengths of graphical regularization and Principal Component Analysis. Comprehensive evaluations on simulated and multi-resolution spatial transcriptomic datasets generated from various platforms demonstrate the capacity of GraphPCA to enhance downstream analysis tasks including spatial domain detection, denoising, and trajectory inference compared to other state-of-the-art methods.

Sparse representation learning for scalable single-cell RNA sequencing data analysis

Kai Zhao, Hon-Cheong So, ♦Zhixiang Lin

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

The rapid rise in the availability and scale of scRNA-seq data needs scalable methods for integrative analysis. Though many methods for data integration have been developed, few focus on understanding the heterogeneous effects of biological conditions across different cell populations in integrative analysis. Our proposed scalable approach, scParser, models the heterogeneous effects from biological conditions, which unveils the key mechanisms by which gene expression contributes to phenotypes. Notably, the extended scParser pinpoints biological processes in cell subpopulations that contribute to disease pathogenesis. scParser achieves favorable performance in cell clustering compared to state-of-the-art methods and has a broad and diverse applicability.

Decoding Gene Functions: Exploring their Significance in Biological Context

♦Ying Zhu

Fudan University

The prevalence of high-throughput technologies has greatly accelerated the discovery of genes with distinct functions and their links to various diseases. Nonetheless, translating these discoveries into mechanistic insights poses a significant challenge. Here, we introduce novel bioinformatics tools tailored to streamline the interpretation of gene functions within their biological context, i.e., in specific tissue and cell types.

Generative Modeling of Single-cell Dynamics with Deep Diffusion Schrödinger Bridge Model

♦Jingsi Ming

East China Normal University

The dynamic evolution of cell fate is a fundamental scientific question in developmental biology, disease modeling, and regenerative medicine. With the rapid advancement of single-cell

RNA sequencing technologies, researchers are now able to observe transcriptional dynamics during cellular differentiation at unprecedented resolution. However, existing methods suffer from certain limitations. Some lack the generative capacity to model cell developmental trajectory, while others fail to establish a fine-grained correspondence between cellular states. To address these challenges, we present SCOPE, a generative framework for single-cell dynamics modeling based on Schrödinger Bridge theory and deep diffusion models. By learning optimal stochastic paths between observed timepoints, our method enables continuous, interpretable, and generative modeling of cell fate dynamics. It also supports interpolation of unobserved intermediate states and simulation of fate shifts under genetic perturbations. Experimental results demonstrate that our approach outperforms state-of-the-art methods in key tasks such as trajectory reconstruction, state prediction, and perturbation simulation, showing strong generalization ability and biological interpretability. This framework provides a novel computational tool for high-resolution mapping of cell fate landscapes and the study of underlying regulatory mechanisms.

25CHI075: Recent Advances in Statistical Machine Learning: Theory and Algorithms

Word-Level Maximum Mean Discrepancy Regularization for Word Embedding

Youqian Gao, ♦Ben Dai

The Chinese University of Hong Kong, The Chinese University of Hong Kong

The word embedding technique is widely utilized in natural language processing (NLP) to represent words as numerical vectors in textual datasets. However, estimating word embeddings may suffer severely from overfitting due to the enormous variety of words. To address this issue, this article proposes a novel regularization framework that explicitly accounts for "word-level distribution discrepancy," a common phenomenon in various NLP tasks where word distributions differ significantly across distinct labels. The proposed regularization, termed word-level maximum mean discrepancy (w-MMD), is a variant of the maximum mean discrepancy (MMD) method specifically designed to enhance and preserve distributional discrepancies within word embedding vectors, thereby preventing overfitting. Our theoretical analysis demonstrates that w-MMD effectively acts as a dimensionality reduction technique for word embeddings, significantly improving the robustness and generalization capabilities of NLP models. The numerical effectiveness of w-MMD and its variants is validated through extensive simulations and empirical studies on the CE-T1 and BBC News datasets, using state-of-the-art deep learning architectures in NLP.

Functional data analysis via neural networks

♦Jun Fan

Hong Kong Baptist University

Neural networks have demonstrated remarkable versatility in approximating continuous functions, but their potential extends even further. In this talk, we explore the domain of functional neural networks, which provide a promising method for approximating nonlinear smooth functionals. By examining the convergence rates of both approximation and generalization errors, we gain insights into the theoretical characteristics of these networks within the empirical risk minimization framework. This investigation enhances our understanding of functional neural networks and paves the way for their effective use in functional data analysis.

Understanding token selection in the self-attention mechanism

Zihao Li, ♦Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason Klusowkski, Jianqing Fan, Mengdi Wang

Princeton University, The University of Hong Kong, Princeton University, Princeton University, Northwestern University, Princeton University, Princeton University, Princeton University

Transformers have emerged as a dominant force in machine learning, showcasing unprecedented success in a wide range of applications. Their unique architecture, characterized by self-attention mechanisms, has revolutionized the way models process data. In this talk, we delve into a series of theoretical case studies focused on understanding token selection within the self-attention mechanism. We first demonstrate that a one-layer transformer model can be successfully trained by gradient descent to perform one-nearest neighbor prediction in context. Then, we show the capacity of one-layer transformers to learn variable selection and solve linear regression with group sparsity. We also investigate the capability of simple transformer models in learning random walks. At the core of these theoretical studies is to analyze how the softmax self-attention can be trained to perform reasonable token selection.

Why Does Differential Privacy Noise Have Limited Impact When Fine-Tuning Large Language Models?

♦Chendi Wang

Xiamen Univeristy

Recent work shows that large-scale pretraining on public datasets significantly improves the performance of differentially private (DP) learning on downstream tasks. We examine this effect through the lens of representation learning, analyzing both the last and intermediate layers of neural networks. For the last layer, we consider a layer-peeled model exhibiting Neural Collapse (NC), showing that misclassification error becomes dimension-independent when actual features are sufficiently close to their ideal forms. Empirically, stronger pretrained models—such as Vision Transformers (ViTs)—offer better last-layer representations, though DP fine-tuning remains more sensitive to perturbations than non-private training. To address this, we adopt feature normalization and PCA, which substantially improve DP fine-tuning accuracy.

We further explore intermediate layers, studying how DP noise affects feature separability in ViTs and representation quality in large language models fine-tuned for reasoning. Using a law of representation learning, we quantify the impact of DP noise across layers and find that, without careful hyperparameter tuning, high privacy budgets degrade representation quality. However, with optimized hyperparameters, DP noise has limited impact, allowing high accuracy under strong privacy guarantees. Our findings demonstrate how public pretraining and principled strategies can mitigate the privacy-utility trade-off in DP deep learning.

25CHI076: Recent Advances on the Analysis of

Failure Time Data

Estimation and Variable Selection for Interval-Censored Failure Time Data with Random Change Point and Application to Breast Cancer Study

♦*Mingyue Du, Yichen Lou, Jianguo Sun*

Jilin University, The Chinese University of Hong Kong, University of Missouri

Motivated by a breast cancer study, we consider regression analysis of interval-censored failure time data in the presence of a random change point. Although a great deal of literature on interval-censored data has been established, there does not seem to exist an established method that can allow for the existence of random change points. Such data can occur in, for example, clinical trials where the risk of a disease may dramatically change when some biological indexes of the human body exceed certain thresholds. To fill the gap, we will first consider regression analysis of such data under a class of linear transformation models and provide a sieve maximum likelihood estimation procedure. Then a penalized method is proposed for simultaneous estimation and variable selection, and the asymptotic properties of the proposed method are established. An extensive simulation study is conducted and indicates that the proposed methods work well in practical situations. The approaches are applied to the real data from the breast cancer study mentioned above. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Linearized maximum rank correlation estimation of doubly truncated data

♦*Peijie Wang, Qihao Wang, Jianguo Sun*

Jilin University, Jilin University, University of Missouri

Truncated data frequently arise in many areas such as economics, astronomical studies and survival analysis. The presence of truncation makes the statistical inference more difficult due to the incomplete information. We proposed a linearized maximum rank correlation estimation of doubly truncated data under the single-index model. Unlike the existing methods, the proposed estimation has a closed-form expression and does not need knowledge of the unknown link function or the error distribution, which makes it more appealing in theory and computation. The estimators are shown to be consistent and asymptotically normal when the linearity of the expectation assumption is satisfied. An extensive simulation study is conducted and indicates that the proposed method works well in various practical situations. The method is further demonstrated by applying it to an AIDS study.

Goodness of fit test for bivariate interval-censored survival data

Bernard Rosner, Camden Bay, Robert Glynn, Gui-Shuan Ying, Maureen Maguire, ♦*Mei-Ling Ting Lee*

Harvard Medical School, Brigham and Women's Hospital, Harvard Medical School, University of Pennsylvania, University of Pennsylvania, University of Maryland

We develop a goodness-of-fit test for general bivariate interval-censored survival data and apply it to the proposed interval-censored version of the bivariate Clayton–Oakes model.

Bayesian estimation of partial functional Tobit censored quantile regression model

♦*Chunjie Wang, Zhexin Lu, Chuchu Wang, Xinyuan Song*

Cahngchun University of Technology, Cahngchun University of Technology, The Chinese University of Hong Kong, The Chinese University of Hong Kong

Imaging data is crucial for disease diagnosis. This study introduces a PFTCQR model to explore relationships between laryngeal cancer onset and imaging/clinical predictors using data from a Chinese hospital. Functional principal component analysis and moment methods estimate predictor functions, while an MCMC algorithm aids estimation via the ALD representation. The model is extended for composite quantile regression with variable selection, improving robustness. Results offer insights into laryngeal cancer risk factors and their quantile-specific effects.

25CHI078: Recent development in statistical methods for related regression models and applications

Sequential quantile regression for streaming data by least squares

♦*Ye Fan, Nan Lin*

Capital University of Economics and Business, Washington University in St. Louis

Massive streaming data are common in modern economics applications, such as e-commerce and finance. They cannot be permanently stored due to storage limitation, and real-time analysis needs to be updated frequently as new data become available. In this paper, we develop a sequential algorithm, SQR, to support efficient quantile regression (QR) analysis for streaming data. Due to the non-smoothness of the check loss, popular gradient-based methods do not directly apply. Our proposed algorithm, partly motivated by the Bayesian QR, converts the non-smooth optimization into a least squares problem and is hence significantly faster than existing algorithms that all require solving a linear programming problem in local processing. We further extend the SQR algorithm to composite quantile regression (CQR), and prove that the SQR estimator is unbiased, asymptotically normal and enjoys a linear convergence rate under mild conditions. We also demonstrate the estimation and inferential performance of SQR through simulation experiments and a real data example on a US used car price data set.

Assessing mediation in cross-sectional stepped wedge cluster randomized trials

♦*Zhiqiang Cao, Fan Li*

Shenzhen Technology University, Yale University

Mediation analysis has been comprehensively studied for independent data but relatively little work has been done for correlated data, especially for the increasingly adopted stepped wedge cluster randomized trials (SW-CRTs). Motivated by challenges in underlying the effect mechanisms in pragmatic and implementation science clinical trials, we develop new methods for mediation analysis in SW-CRTs. Specifically, based on a linear and generalized linear mixed models, we demonstrate how to estimate the natural indirect effect and mediation proportion in typical SW-CRTs with four data types, including both continuous and binary mediators and outcomes. Furthermore, to address the emerging challenges in exposure-time treatment effect heterogeneity, we derive the mediation expressions in SW-CRTs when the total effect varies as a function of the exposure time. The cluster jackknife approach is considered for inference across all data types and treatment effect structures. We conduct extensive simulations to evaluate the finite-sample performances of proposed mediation estimators and demonstrate the proposed approach in a real data example. A user-friendly R package mediateSWCRT has been developed to facilitate the practical implementation of the estimators.

Nonparametric Sensitivity Analysis for Unobserved Confounding with Survival Outcomes

♦Rui Hu, Ted Westling

Shenzhen Technology University, University of Massachusetts Amherst

In observational studies, the observed association between an exposure and outcome of interest may be distorted by unobserved confounding. Causal sensitivity analysis is often used to assess the robustness of observed associations to potential unobserved confounding. For time-to-event outcomes, existing sensitivity analysis methods rely on parametric assumptions on the structure of the unobserved confounders and Cox proportional hazards models for the outcome regression. If these assumptions fail to hold, it is unclear whether the conclusions of the sensitivity analysis remain valid. Additionally, causal interpretation of the hazard ratio is challenging. To address these limitations, in this paper we develop a nonparametric sensitivity analysis framework for time-to-event data. Specifically, we derive nonparametric bounds for the difference between the observed and counterfactual survival curves and propose estimators and inference for these bounds using semiparametric efficiency theory. We also provide nonparametric bounds and inference for the difference between the observed and counterfactual restricted mean survival times. We demonstrate the performance of our proposed methods using numerical studies and an analysis of the causal effect of physical activity on respiratory disease mortality among former smokers.

Matrix Autoregressive Time Series with Reduced-Rank and Sparse Structural Constraints

♦Xiaohang Wang, Ling Xin, Philip L.H. Yu

The Education University of Hong Kong, Beijing Normal-Hong Kong Baptist University, The Education University of Hong Kong

Matrix- and tensor-valued time series models have been explored as one of the ways to relieve the dimensionality issue in high-dimensional time series modelling. It uses the multi-classification structures in data variables to decompose large interaction networks into smaller ones. With the need for further dimension reduction, recent studies have considered imposing further structural constraints on the individual coefficient matrix of the matrix- or tensor-valued time series models. In this paper, we propose the RR-S-MAR model, which is a matrix autoregressive (MAR) model with a reduced-rank structure on the left matrices and a sparse structure on the right matrices. We address the estimation, inference, selection and interpretation issues of the proposed model. We develop an alternating least-squares method to estimate the constrained model and adopt bootstrapping method for statistical inference. An extended Bayesian information criterion is proposed for selecting the tuning parameters in the model. Simulations are used to study the performance of the estimation algorithm and the model selection criterion in finite samples. An economic data is used to demonstrate the real analysis and model interpretations.

25CHI079: Recent developments about high-dimensional inference

Targeted Inference for High-Dimensional Quantile Regression Models

♦Xuejun Jiang, Yakun Liang

Department of Statistics and Data Science, Southern University of Science and Technology, Department of Statistics and Data Science, Southern University of Science and Technology

This research introduces an innovative inference framework that employs dimension reduced convolution-smoothed quantile regression, while avoiding estimating the inverse of high-dimensional covariance matrix of the predictors. By calibrating the regularization parameter, we develop a data-driven test that can be shown to be an oracle test with probability tending to one. To mitigate the selective bias induced by dimension reduction and ensure valid inference, we implement a cross-fitting strategy by dividing the dataset into two parts: one for model selection and the other for parameter estimation. This process yields a fused estimator, derived from an informative weighting method that combines estimators from both dataset partitions. The fused estimator aids in constructing confidence intervals and performing Wald-type tests for targeted parameters. We establish the Bahadur representation of this estimator and obtain limiting distributions of the test statistics under both null and alternative hypotheses, with the number of parameters diverging to infinity. Advantages of our tests are further highlighted by theoretical power comparisons to some competitive tests. Empirical studies confirm effectiveness of the proposed tests across various linear parameter hypotheses. Additionally, we illustrate the use of the proposed methodology through two real-world data analyses.

Robust Mutual Fund Selection with False Discovery Rate Control

Hongfei Wang, Ping Zhao, ♦Long Feng, Zhaojun Wang

Nanjing Audit University, Nankai University, Nankai University, Nankai University

In this article, we address the challenge of identifying skilled mutual funds among a large pool of candidates, utilizing the linear factor pricing model. Assuming observable factors with a weak correlation structure for the idiosyncratic error, we propose a spatial-sign based multiple testing procedure (SS-BH). When latent factors are present, we first extract them using the elliptical principle component method (He et al. 2022) and then propose a factor-adjusted spatial-sign based multiple testing procedure (FSS-BH). Simulation studies demonstrate that our proposed FSS-BH procedure performs exceptionally well across various applications and exhibits robustness to variations in the covariance structure and the distribution of the error term. Additionally, real data application further highlights the superiority of the FSS-BH procedure.

Homogeneity tests of high-dimensional covariance matrices with applications to change-points detection

Jiayu Lai, Xiaoyi Wang, ♦Le Zhou, Shurong Zheng

Northeast Normal University, Beijing Normal University, Hong

Kong Baptist University, Northeast Normal University

In this paper, we built a series of tests based on U-statistics for testing the high-dimensional covariance matrix change-points. The asymptotic distributions of the constructed U-statistics are derived under the null and local alternative hypotheses. Then, we propose a family of maximum-type statistics, after which two test methods based on the combination of the p-values of these maximum-type statistics are developed. We also propose three new adaptive methods to estimate the location of the change-point and obtain the corresponding convergence rate. Finally, the binary segmentation method is proposed to be combined with our three adaptive estimators to detect multiple change-points. Our simulation study shows that the proposed test methods can maintain high powers under alternatives with different sparsity levels and that our proposed adaptive estimators perform well under different alternatives with both single and multiple change-points.

Testing the general linear hypothesis in high-dimensional heteroscedastic factor model via random integration

♦*Mingxiang Cao*

Anhui Normal University

In this paper, we proposed a test procedure based on random integration and the well-known Welch–Satterthwaite chi-square-approximate method to deal with the problem of general linear hypothesis testing (GLHT) in high-dimensional heteroskedastic factor model. The asymptotic distributions of the proposed test statistic were obtained under the null and the alternative hypotheses, respectively. The results showed that it was more reasonable to approximate the distributions of the new test using the distribution of the chi-square type mixture. Numerical simulations and real data analysis further showed that our proposed test was more powerful than competing tests.

25CHI081: Recent developments in causal inference and survival analysis

SEMIPARAMETRIC CURE REGRESSION MODELS WITH INFORMATIVE CASE K INTERVAL-CENSORED FAILURE TIME DATA

♦*Yichen Lou, Jianguo Sun, Peijie Wang*

The Chinese University of Hong Kong, University of Missouri, Jilin University

Interval-censored failure time data occur in many areas and many methods for their analyses have been proposed. In particular, some methods have been developed for the situation with the existence of a cured subgroup or informative censoring. In this paper, we discuss the case where both a cured subgroup and informative censoring exist and a frailty-based semiparametric non-mixture cure model approach is proposed. For inference, a two-step estimation procedure is developed and the resulting estimators of regression parameters are shown to be consistent and asymptotically normal. An extensive simulation study is conducted and indicates that the proposed procedure works well in practice. In addition, the methodology is applied to a set of real data arising from an Alzheimer's disease study.

Significance test for semiparametric conditional average treatment effects and other structural functions

♦*Niwen Zhou, Xu Guo, Lixing Zhu*

Beijing Normal University, Beijing Normal University, Beijing Normal University

The paper investigates a hypothesis testing problem concerning the potential additional contributions of other covariates to the structural function, given the known covariates.

The structural function is the conditional expectation given covariates in which the

response may depend on unknown nuisance functions. It includes classic regression functions and the conditional average treatment effects as illustrative instances. Based on Neyman's orthogonality condition, the proposed distance-based test exhibits the quasi-oracle property in the sense that the nuisance function asymptotically does not influence on the limiting distributions of the test statistic under both the null and alternatives. This novel test can effectively detect the local alternatives distinct from the null at the fastest possible rate in hypothesis testing. This is particularly noteworthy given the involvement of nonparametric estimation of the conditional expectation. Numerical studies are conducted to examine the performance of the test. In the real data analysis section, the proposed tests are applied to identify significantly explanatory covariates that are associated with AIDS treatment effects, yielding noteworthy insights.

Causal inference for time-to-event data with a cured subpopulation

♦*Yi Wang, Yuhao Deng, Xiaohua Zhou*

Shanghai University of International Business and Economics, Peking University, Peking University

When studying the treatment effect on time-to-event outcomes, it is common that some individuals never experience failure events, which suggests that they have been cured. However, the cure status may not be observed due to censoring, which makes it challenging to define treatment effects. Current methods mainly focus on estimating model parameters in various cure models, ultimately leading to a lack of causal interpretations. To address this issue, two causal estimands are proposed, the timewise risk difference and mean survival time difference, in the always-uncured based on principal stratification as a complement to the treatment effect on cure rates. These estimands allow the study of the treatment effects on failure times in the always-uncured subpopulation. The identifiability is shown using a substitutional variable for the potential cure status under the ignorable treatment assignment mechanism; these two estimands are identifiable. Estimation methods are also provided using mixture cure models. The approach is applied to an observational study that compared the leukemia-free survival rates of different transplantation types to cure acute lymphoblastic leukemia. The proposed approach yielded insightful results that can be used to inform future treatment decisions.

Recent developments in causal inference and survival analysis

Chang Wang, ♦*Baihua He, Shishun Zhao, Jianguo Sun, Xinyu Zhang*

University of Science and Technology of China

A large literature has been established for random survival forest (RSF), a popular tool developed to analyze right-censored failure time data, under various situations. However, its prediction performance may not be optimal sometimes. To address this issue, we propose two optimal model averaging methods based on martingale residual processes. In particular, an in-of-bag and out-of-bag (IBOB) data process is defined, and two new IBOB functionals criteria are derived for the selection of weights. Furthermore, for their implementation, a greedy algorithm is presented, and the asymptotic optimality of the proposed model averaging approaches is established along with the convergence of the greedy averaging algorithms. Finally, an extensive simulation study is conducted, which indicates that the proposed methods work well, and an illustration is provided.

25CHI082: Recent Developments in Covariate Adjustment for Randomized Clinical Trials

Bias Reduction in G-computation for Covariate Adjustment in Randomized Clinical Trials

♦Xin Zhang, Lin Liu, Haitao Chu

Pfizer Inc, Shanghai Jiao Tong University, Pfizer Inc and University of Minnesota

G-computation has become a widely used method for estimating unconditional treatment effects with covariate adjustment in the analysis of randomized clinical trials. It relies on fitting canonical generalized linear models (GLMs), which could be problematic with sparse data or rare events. Firth correction (FC) has become a popular approach in practice to address those issues when fitting GLMs. A naive use of FC, instead of maximum likelihood estimation (MLE), with g-computation is statistical valid for estimation and inference as justified by M-estimation theory, however, it amplifies the bias and harms the efficiency at finite sample sizes. This demands a thorough study of the asymptotic bias of g-computation estimators to improve the use of FC. In this work, we propose a novel approach to the reduction of the asymptotic bias of g-computation for both estimation and inference. The proposed approach is developed under simple randomization and without assuming correct specification of working models. Those resulting bias-reduced estimators are convenient to implement via generalized Oaxaca-Blinder estimators, equipped with the estimated nuisance parameters obtained from minor modification of MLE/FC estimators. Through extensive simulations, we demonstrate the superior finite-sample performance of the proposed method.

Robust and Efficient Statistical Inference Under Covariate-Adaptive Randomization

♦Fuyi Tu

School of Science, Chongqing University of Posts and Telecommunications

Randomized controlled trials have been considered as the gold standard in clinical research. However, covariate-adaptive randomization, which enhances trial efficiency by balancing covariates across treatment groups, may introduce correlations between treatment assignments and outcomes, thereby invalidating conventional statistical inference. To address this, we propose a robust and efficient framework for covariate-adaptive randomization, employing linear regression as a working model without specifying the form between covariates and outcomes. In the presence of high-dimensional covariates, the ordinary-least-squares estimators tend to fail due to over-fitting, we then develop a more general framework applicable to both low- and high-dimensional settings, and validate two Lasso-adjusted estimators for high-dimensional cases. When there is no strong evidence of a linear relationship, nonparametric or machine learning approaches can be utilized. The consistency of our proposed treatment effect estimators and variance estimators is theoretically assured and empirically validated through numerical simulations and real-world data analyses. Finally, we provide practical recommendations for selecting treatment effect estimators and statistical inference methods based on simplicity and efficiency.

Covariate Adjustment in Randomized Experiments Motivated by Higher-Order Influence Functions

♦Lin Liu

Shanghai Jiao Tong University

In this talk, we will demonstrate that the recently proposed covariate adjustment methodologies can be viewed as special cases of the higher-order influence function (HOIF) approaches, which are designed for constructing rate-optimal estimators for statistical functionals under minimal complexity-reducing assumptions. We will discuss the application of HOIFs in several settings, including randomization-based inference and beyond.

Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates

Xin Lu, Fan Yang, ♦Yuhao Wang

Tsinghua University, Tsinghua University, Tsinghua University

Completely randomized experiment is the gold standard for causal inference. When the covariate information for each experimental candidate is available, one typical way is to include them in covariate adjustments for more accurate treatment effect estimation. In this paper, we investigate this problem under the randomization-based framework, i.e., that the covariates and potential outcomes of all experimental candidates are assumed as deterministic quantities and the randomness comes solely from the treatment assignment mechanism. Under this framework, to achieve asymptotically valid inference, existing estimators usually require either (i) that the dimension of covariates p grows at a rate no faster than O(n^{3/4}) as sample size n \to \infinity; or (ii) certain sparsity constraints on the linear representations of potential outcomes constructed via possibly high-dimensional covariates. In this paper, we consider the moderately high-dimensional regime where p is allowed to be in the same order of magnitude as n. We develop a novel debiased estimator with a corresponding inference procedure and establish its asymptotic normality under mild assumptions. Our estimator is model-free and does not require any sparsity constraint on potential outcome's linear representations. We also discuss its asymptotic efficiency improvements over the unadjusted treatment effect estimator under different dimensionality constraints. Numerical analysis confirms that compared to other regression adjustment based treatment effect estimators, our debiased estimator performs well in moderately high dimensions.

25CHI083: Recent Developments in High-dimensional Data Analysis

Integrative Analysis of High-dimensional RCT and RWD Subject to Censoring and Hidden Confounding

♦Xin Ye, Shu Yang, Xiaofei Wang, Yanyan Liu

School of Statistics and Mathematics, Guangdong University of Finance and Economics, Department of Statistics, North Carolina State University, Department of Biostatistics and Bioinformatics, Duke University, School of Mathematics and Statistics, Wuhan University

In this study, we focus on estimating the heterogeneous treatment effect (HTE) for survival outcomes. The outcome is subject to censoring and the number of covariates is high-dimensional. We utilize data from both the randomized controlled trial (RCT), considered as the gold standard, and real-world data (RWD), which may be affected by hidden confounding factors. To achieve a more efficient HTE estimate, such integrative analysis requires deep insight into the data generation mechanism, particularly the accurate characterization of unmeasured confounding effects/bias. With this aim, we propose a penalized-regression-based integrative approach that allows for the simultaneous estimation of parameters, selection of variables, and identification of the existence of unmeasured confounding effects. The consistency, asymptotic normality, and efficiency gains are rigorously established for the proposed estimate. Finally, we apply the proposed method to estimate the HTE of lobar/sublobar resection on the survival of lung cancer patients. The RCT is a multicenter non-inferiority randomized phase 3 trial, and the RWD comes from a clinical oncology cancer registry in the United States. The analysis reveals that the unmeasured confounding exists, and the integrative approach does enhance the efficiency for the HTE estimation.

Kernel Variable Importance Measure with Applications

♦*Bingyao Huang, Guanghui Cheng, Yanyan Liu, Liuhua Peng*

Guangdong University of Technology, Guangzhou University, Wuhan University, The University of Melbourne

This paper introduces a novel kernel variable importance measure (KvIM) based on the maximum mean discrepancy (MMD). KvIM can effectively measure the importance of each individual dimension in contributing to the distributional difference by constructing weighted MMD and applying perturbations to evaluate changes in MMD through assigned weights. KvIM has several notable advantages: it is nonparametric and model-free, accounts for dependencies among dimensions, and is suitable for high-dimensional data. We establish the consistency of the empirical KvIM under general conditions, along with its theoretical properties in high-dimensional settings. Furthermore, we apply KvIM to classification problems and streaming datasets, proposing a KvIM-enhanced classification approach and an online KvIM. These applications demonstrate the practical utility of the proposed KvIM in diverse scenarios, as justified through extensive numerical experiments.

Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

♦*Xiangbin Hu*

Beijing Institute of Technology

We propose a provable deep conditional generative framework for data-driven joint pricing and inventory control with contextual information, in which the retailer makes pricing and inventory decision based solely on historical data consisting of demand, price and covariates, In contrast to the existing methods that directly learn the optimal policy curve as a function of the covariate, we develop a deep conditional generative learning approach to learn from historical data the whole demand distribution conditional on selling price and covariates. The conditional demand estimator is a deep conditional generator, which is a function of the selling price, covariates and a 'noise' variable independent of price and covariates. The special structure of the deep conditional generator allows us to sample the noise easily and compute the gradient of the random cost in closed form, with which stochastic gradient descent can be designed for the subsequent policy optimization under both risk neutral and risk averse settings. We establish asymptotic properties of our data-driven policy including asymptotic optimality and consistency, and we provide a consistent estimator of the optimal expected cost. Through comprehensive simulations, we show that the expected cost evaluated at our data-driven policy is significantly smaller than those from state-of-the-art prescriptive benchmarks. In particular, we apply our method to real meal delivery data and find that our method beats the benchmarks by at least 10% reduction in out-of-sample costs, and it is at least four times faster than three machine learning-based descriptive approaches.

Non-parametric inference based on reliability life-test of non-identical coherent systems

♦*Xiaojun Zhu*

Xi'an Jiaotong-Liverpool University

The talk discusses non-parametric and semi-parametric statistical inferential methods for estimating component and system lifetimes based on life-tests of non-identical coherent systems with known signatures. The estimations are obtained through maximum likelihood estimation method upon using EM-algorithm and the cumulative hazard approach of Nelson–Aalen estimation method. They are extended to multiple-stress problems and can be used to test the impact of system structure on lifetimes.

25CHI084: Recent developments in reinforcement learning and mobile health

Factorial Causal Excursion Effects: Modeling Time-Varying Effects of Multi-Component Mobile Health Interventions in Micro-Randomized Trials

♦*Xueqing Liu, Weihao Li, Bibhas Chakraborty*

Duke-NUS Medical School, National University of Singapore, Duke-NUS Medical School

Recent advances in mobile health technologies have spurred the development of just-in-time adaptive interventions (JITAIs), which dynamically tailor treatments to an individual's evolving needs and contexts, delivering the right intervention at the right moment. Micro-randomized trials (MRTs) are critical in designing and evaluating JITAIs, wherein which participants are sequentially randomized among intervention options across hundreds or even thousands of decision points. In many practical situations, researchers must evaluate multiple intervention components simultaneously. This need has driven the adoption of factorial MRTs, where multiple components are randomized concurrently at each decision point, all aimed at the same proximal outcome. However, existing causal inference methods for MRT data primarily address single-component interventions, limiting their applicability to complex, multi-component scenarios.

Motivated by this challenge, we introduce the concept of factorial causal excursion effects (CEEs), including both conditional and general CEEs, which quantify the main effect of each treatment while keeping the other treatments at a baseline or marginalizing over them. We propose two estimators, one for each type of CEE, which employ cross-fitting techniques and are robust to outcome model misspecification. We establish their asymptotic properties, validate their performance through simulations, and apply our methods to the DIAMANTE MRT data to assess the impact of text messages on short-term physical activity.

Minimax Regret Learning for Data with Heterogeneous Sub-populations

♦Weibin Mo, Weijing Tang, Songkai Xue, Yufeng Liu, Ji Zhu

Purdue University, Carnegie Mellon University, University of Michigan, University of North Carolina, Chapel Hill, University of Michigan

Modern complex datasets often consist of various sub-populations. To develop robust and generalizable methods in the presence of sub-population heterogeneity, it is important to guarantee a uniform learning performance instead of an average one. In many applications, prior information is often available on which sub-population or group the data points belong to. Given the observed groups of data, we develop a min-max-regret (MMR) learning framework for general supervised learning, which targets to minimize the worst-group regret. Motivated from the regret-based decision theoretic framework, the proposed MMR is distinguished from the value-based or risk-based robust learning methods in the existing literature. The regret criterion features several robustness and invariance properties simultaneously. In terms of generalizability, we develop the theoretical guarantee for the worst-case regret over a super-population of the meta data, which incorporates the observed sub-populations, their mixtures, as well as other unseen sub-populations that could be approximated by the observed ones. We demonstrate the effectiveness of our method through extensive simulation studies and an application to kidney transplantation data from hundreds of transplant centers.

A Synergetic Random Forest Framework for Policy Evaluation

Rui Qiu, Zexuan Zhang, Zhou Yu, ♦Ruoqing Zhu

Beiing University, University of Illinois Urbana Champaign, East China Normal University, University of Illinois Urbana Champaign

We propose a new breed of random forest model in which the splitting rule depends not only on the within-node data, but also information across the entire tree. The new model is particularly suited for situations when the splitting rule, while viewed as an estimating equation, requires further estimation of nuisance parameters that are not feasible within the node. In our proposed model, the nuisance parameter estimation is synergized across the entire tree and also progressively grows as tree nodes expand, facilitating the estimation of the main parameter of interest. A typical use case of such a model is policy evaluation in reinforcement learning, when estimating the value function can utilize information of the transitional kernel. Utilizing the platform of random forests, we can also easily quantify the uncertainty of policy evaluation, which can often be challenging with other approaches.

Online statistical inference for robust policy evaluation in reinforcement learning

Weidong Liu, Jiyuan Tu, Xi Chen, ♦Yichen Zhang

SJTU, SUFE, New York University, Purdue University

Reinforcement learning has recently emerged as a central topic in modern statistics, with policy evaluation playing a fundamental role. Departing from the conventional machine learning literature, which often emphasizes algorithmic performance, our work focuses on statistical inference for the parameters estimated by reinforcement learning algorithms. Existing analyses typically assume that rewards follow standard distributions, which can limit their robustness and applicability. In contrast, we adopt a robust statistical perspective by addressing both outlier contamination and heavy-tailed rewards within a unified framework. We propose an online robust policy evaluation method and derive the limiting distribution of the resulting estimator via its Bahadur representation. Building on this, we further develop a fully-online procedure for efficient statistical inference using the asymptotic distribution. By integrating robust statistics with rigorous inferential tools, this work provides a more reliable and adaptable framework for policy evaluation in reinforcement learning. We demonstrate the practical effectiveness of our approach through numerical studies on real-world reinforcement learning tasks.

25CHI088: Robust Analysis for Treatment Decision and Risk Prediction under Complex Data Settings

Semiparametric Regression Analysis for Interval-Censored Outcome Subject to Misdiagnosis

♦Yuhao Deng, Donglin Zeng, Yuanjia Wang

University of Michigan, University of Michigan, Columbia University

Interval-censored data arise when a disease event is diagnosed at intermittent follow-up times. The disease diagnosis is usually assumed to be accurate in most methods for analyzing interval-censored data. However, this may not be true in practice due to inaccurate measurements of diagnosis biomarkers or clinical knowledge limitations. In this work, we study semiparametric regression models for analyzing interval-censored data assuming imperfect diagnosis at each follow-up time. The model consideration allows death as a competing risk and can incorporate post-mortem autopsy data for disease diagnosis. For inference, we propose a nonparametric maximum likelihood approach for estimation via EM algorithms. We show that the obtained estimators are asymptotically normal and achieve semiparametric efficiency bound. Application to the ADNI data is used to illustrate our approach.

INFERENCE FOR HIGH DIMENSIONAL PROPORTIONAL HAZARDS MODEL WITH STREAMING SURVIVAL DATA

♦Dongxiao Han

School of Statistics and Data Science, Nankai University

We propose an online inference procedure for high dimensional streaming survival data based on the proportional hazards model. Specifically, we offer an online Lasso method for regression parameter estimation and establish the non-asymptotic error bounds of the corresponding Lasso estimators for the regression parameter vector. In addition, we study the pointwise inference for the regression parameters by utilizing a debiased Lasso approach. Furthermore, we conduct high dimensional group inference for the regression parameters based on quadratic forms of the Lasso estimators and the debiased Lasso idea. Extensive simulations are conducted to evaluate the finite sample performance of the proposed method. An application to a colon cancer dataset is provided to demonstrate the practical utility of the proposed methodology

A Robust Covariate-Balancing Method for Estimating Individualized Treatment with Censored Data

Rujia Zheng, ♦Wensheng Zhu, Xiaofan Guo

Northeast Normal University, Yunnan University, The First Hospital of China Medical University

One of the most essential aspects of precision medicine is the identification of optimal individualized treatment regimen, which recommends treatment decisions to maximize patient's expected survival time based on their individual characteristics with censored data. Typically, the expected survival time is required to be estimated first, which is usually based on the posited weighting models (propensity score model and censoring model) or the posited outcome model. However, if any of the above models is misspecified, the estimated treatment regimen is not reliable. In this article, we consider the contrast value function defined for survival analysis, and propose two robust covariate-balancing estimators of the contrast value function by balancing the covariates of patients through censoring probability and survival function of censoring time in the weights, respectively. Theoretical results prove that the proposed estimators are doubly robust, that is, they are consistent if either the propensity score model and the censoring model are correctly specified simultaneously or the outcome model is correctly specified. The asymptotic normality of the estimators is also established under standard regularity. A large number of simulations show the superiority of our methods over the existing methods. Application of the proposed methods is illustrated through analysis of data from the China Rural Hypertension Control Project (CRHCP).

Learning Optimal Early Decision Treatment Rules with Multi-domain Intermediate Outcomes

♦Yuanjia Wang

Implementing precision medicine for mental disorders presents challenges due to disease complexity and heterogeneity in patient responses.Empirical studies suggest that early indicators, such as interim measures (e.g., interim patient selfreports) of disease improvement or relapse, can predict longer-term outcomes, serving as proxies when final outcomes (e.g., in-clinic assessments) are less accessible.

However, existing approaches for deriving individualized treatment rules (ITRs) often ignore these early signals, instead focusing only on a final outcome as the reward. In this work, we propose a new method incorporating intermediate outcomes from various domains into a personalized composite outcome, serving as the reward for learning ITRs.This composite is a weighted sum of inferred latent states from observed measures, with weights personalized for each patient, ensuring consistency with the long-term final response.

Our simulations show that this approach not only provides early detection of non-responders but also improves long-term treatment outcomes.Applying our framework to a randomized clinical trial on major depressive disorder (MDD) demonstrates its effectiveness and advantages in ITR learning.

25CHI093: Some recent developments about big data analysis

Communication-Efficient and Distributed-Oracle Estimation for High-Dimensional Quantile Regression

♦Songshan Yang, Yifan Gu, Hanfang Yang, Xuming He

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Center for Applied Statistics and School of Statistics, Renmin University of China, Center for Applied Statistics and School of Statistics, Renmin University of China, Department of Statistics and Data Science, Washington University in St. Louis

In this article, we present a novel communication-efficient estimator for distributed high-dimensional quantile regression with folded-concave penalties. An iterative multi-step (IM) algorithm is employed to tackle the nonconvex challenge of the objective function, taking into account both the statistical accuracy and the communication constraints. We demonstrate that the proposed IM estimators share similar properties with the global folded-concave penalized estimator. To establish the theoretical results, we introduce a new concept called distributed-oracle estimator. We prove that the proposed estimator converges to the distributed-oracle estimator with high probability. Compared to the L1-penalized method, the IM estimator possesses a faster rate of convergence and requires milder conditions to achieve support recovery. Furthermore, we extend our framework to facilitate distributed inference for the preconceived low-dimensional components within the high-dimensional model. We derive the limiting distribution of the corresponding test statistic under the null hypothesis and the local alternatives. In addition, a new feature-splitting algorithm is devised to accommodate the high-dimensional data within the distributed system. Extensive numerical studies demonstrate the effectiveness and validity of our proposed estimation and inference methods. A real example is also presented for illustration.

Optimal subsampling for high-dimensional partially linear models via machine learning methods

Yujing Shao, ♦Lei Wang, Heng Lian, Haiying Wang

Nankai University

In this paper, we explore optimal subsampling strategies for estimating the parametric regression coefficients in partially linear models with

unknown nuisance functions involving high-dimensional and potentially endogenous covariates. To address model misspecifications and the curse of dimensionality, we leverage flexible machine learning (ML) techniques to estimate the unknown nuisance functions. By constructing an unbiased subsampling Neyman-orthogonal score function, we eliminate regularization bias. A two-step algorithm is then used to obtain appropriate ML estimators of the nuisance functions, mitigating the risk of over-fitting. Using martingale techniques, we establish the unconditional consistency and asymptotic normality of the subsample estimators. Furthermore, we derive optimal subsampling probabilities, including A-optimal and L-optimal probabilities as special cases. The proposed optimal subsampling approach is extended to partially linear instrumental variable models to account for potential endogeneity through instrumental variables. Simulation studies and an empirical analysis of the

Physicochemical Properties of Protein Tertiary Structure dataset demonstrate the superior performance of our subsample estimators.

Least Squares and Hypothesis Testing based Transfer Learning for High-Dimensional Quantile Regression

♦Kangning Wang, Xiaotong Zhu

Shandong Technology and Business University, Shandong Technology and Business University

High-dimensional quantile regression concerns on learning the conditional quantiles based on a specific high-dimensional target data. In real applications, the target sample size is usually too limited to provide an accurate results, while possibly related source datasets are available to make improvements. Then transfer learning plays an important role in such case, this paper proposes least squares and hypothesis testing based transfer learning for high-dimensional quantile regression. More specifically, when the informative set is known, we construct a LASSO least squares based quantile regression transfer learning framework, and establish the estimation error bounds, which are lower than those with target data only. Besides, a hypothesis testing based source detection algorithm is proposed, and we prove that the probability of incorrectly excluding transferable source datasets, i.e., type \uppercase\expandafter{\romannumeral1} error, will become small as the source data size increases. Moreover, the convenient LARS algorithm can be applied to reduce computational complexity. The numerical results confirm the effectiveness of the proposed methods.

25CHI096: Statistical Advances in Large Language Models and Network Analysis

Learning nonparametric graphical model on heterogeneous network-linked data

♦Junhui Wang

Chinese University of Hong Kong

Graphical models have been popularly used for capturing conditional independence structure in multivariate data, which are often built upon independent and identically distributed observations, limiting their applicability to complex datasets such as network-linked data. In this talk, we introduce a nonparametric graphical model that addresses these limitations by accommodating heterogeneous graph structures without imposing any specific distributional assumptions. The introduced estimation method effectively integrates network embedding with nonparametric graphical model estimation. It further transforms the graph learning task into solving a finite-dimensional linear equation system by leveraging the properties of vector-valued reproducing kernel Hilbert space. We will also discuss theoretical properties of the proposed method in terms of the estimation consistency and exact recovery of the heterogeneous graph structures. Its effectiveness is also demonstrated through a variety of simulated examples and a real application to the statistician coauthorship dataset.

A Statistical Hypothesis Testing Framework for Data Misappropriation Detection in Large Language Models

Yinpeng Cai, Lexin Li, ♦Linjun Zhang

Peking University, UC Berkeley, Rutgers University

Large Language Models (LLMs) are rapidly gaining enormous popularity in recent years. However, the training of LLMs has raised significant privacy and legal concerns, particularly regarding the inclusion of copyrighted materials in their training data without proper attribution or licensing, which falls under the broader issue of data misappropriation. In this article, we focus on a specific problem of data misappropriation detection, namely, to determine whether a given LLM has incorporated data generated by another LLM. To address this issue, we propose embedding watermarks into the copyrighted training data and formulating the detection of data misappropriation as a hypothesis testing problem. We develop a general statistical testing framework, construct a pivotal statistic, determine the optimal rejection threshold, and explicitly control the type I and type II errors. Furthermore, we establish the asymptotic optimality properties of the proposed tests, and demonstrate its empirical effectiveness through intensive numerical experiments.

Robust Reinforcement Learning from Human Feedback for Large Language Models Fine-Tuning

Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, ♦Chengchun Shi

LSE, Tsinghua University, LSE, Oxford, London School of Economics and Political Science

most existing RLHF algorithms use the Bradley-Terry model, which relies on assumptions about human preferences that may not reflect the complexity and variability of real-world judgments. In this paper, we propose a robust algorithm to enhance the performance of existing approaches under such reward model misspecifications. Theoretically, our algorithm reduces the variance of reward and policy estimators, leading to improved regret bounds. Empirical evaluations on LLM benchmark datasets demonstrate that the proposed algorithm consistently outperforms existing methods, with 77-81 % of responses being favored over baselines on the Anthropic Helpful and Harmless dataset.

A Statistical Take on Watermarks for Large Language Models: Theory, Applications, and Future Opportunities

♦Weijie Su

University of Pennsylvania

Watermarking, the embedding of subtle statistical signals into text generated by large language models (LLMs), has become a principled approach to distinguishing synthetic text from human-written content. In this talk, we provide a statistical overview of watermarking techniques for LLMs, focusing on their theoretical foundations and practical applications. First, we introduce a general framework for evaluating the statistical efficiency of watermarks, which enables the design of optimal detection rules with rigorous control of false positive and false negative rates. Building on this framework, we address the challenge of robust detection under extensive human edits, proposing a truncated goodness-of-fit approach that adaptively achieves optimal detection power in mixture-model settings, outperforming conventional sum-based methods. We then consider practical scenarios involving mixed-source text, blending human-written and watermarked content, and demonstrate how to estimate the proportion of watermarked text for attribution purposes. Empirical results across diverse scenarios validate the effectiveness of these methods. We conclude the talk by exploring several future directions and open

questions in watermarking research.

25CHI098: Statistical analysis of complex survival data

Semiparametric Analysis of Additive–Multiplicative Hazards Model with Interval-Censored Data and Panel Count Data

Tong Wang, Yang Li, Jianguo Sun, ♦Shuying Wang

Changchun University of Technology, Changchun University of Technology, University of Missouri, Changchun University of Technology

In survival analysis, interval-censored data and panel count data represent two prevalent types of incomplete data. Given that, within certain research contexts, the events of interest may simultaneously involve both data types, it is imperative to perform a joint analysis of these data to fully comprehend the occurrence process of the events being studied. In this paper, a novel semiparametric joint regression analysis framework is proposed for the analysis of interval-censored data and panel count data. It is hypothesized that the failure time follows an additive–multiplicative hazards model, while the recurrent events follow a nonhomogeneous Poisson process. Additionally, a gamma-distributed frailty is introduced to describe the correlation between the failure time and the count process of recurrent events. To estimate the model parameters, a sieve maximum likelihood estimation method based on Bernstein polynomials is proposed. The performance of this estimation method under finite sample conditions is evaluated through a series of simulation studies, and an empirical study is illustrated.

Group penalized doubly nonparametric probit model with interval censored survival data and application to pharyngeal disease

♦Bo Zhao, Chunjie Wang, Shuying Wang, Dan Yu

Changchun University of Technology, Changchun University of Technology, Changchun University of Technology, Department of Otolaryngology Head and Neck Surgery the Second Hospital, Jilin University

Interval-censored failure time data frequently arise in many fields such as medical studies, demography, economics and social sciences, and the main feature is that the failure time is not observed exactly, but is known to belong to a window or an interval. The existing analysis method is to establish linear regression analysis including semiparametric additive hazards and proportional hazards models. In this paper, we propose a doubly nonparametric probit model by combining the transformation function of failure time and nonparametric additive function regarding covariates that can be decomposed into multiple additive univariate functions. And the number of variables and additive components may be larger than the sample size but the number of nonzero additive components is "small" relative to the sample size. A very important issue is to determine the nonzero additive components. To deal with this problem, we propose a sieve group penalized variable selection procedure that involves minimizing a negative sieve log-likelihood function plus a group penalization, in which the Bernstein polynomials splines are used to approximate transformation functions and nonparametric functions. Furthermore, a computationally efficient group coordinate descent algorithm is developed to implement the proposed method. Extensive numerical simulation studies indicate that the proposed method works well, and also report that the proposed method is robust to the assumption that the true model of the failure time was misspecified. The proposed method is applied to a pharyngeal disease study for identifying important and relevant clinical factors.

Distributed Least Product Relative Error estimation for semi-parametric multiplicative regression with massive data

♦Xiaohui Yuan

Changchun University

Distributed systems have been widely used for massive data analysis, but few studies focus on multiplicative regression models. We consider a communication-efficient surrogate likelihood method using the Least Product Relative Error criterion for semi-parametric multiplicative models on massive datasets. The non-parametric component is efficiently handled via B-spline approximation. We derive the asymptotic properties for both parametric and non-parametric components, while the SCAD and adaptive Lasso penalty functions are developed and their oracle properties for variable selection are validated. Simulation studies and an application to an energy prediction dataset are used to demonstrate the effectiveness and practical utility of the proposed method.

Bayesian empirical likelihood for accelerated failure time model with covariates missing at random

♦Xinrui Liu

Changchun University of Technology

This study proposes a Bayesian empirical likelihood (BEL) procedure based on the inverse probability weighted (IPW) Buckley-James estimation equation to analyze semiparametric accelerated failure time (AFT) models in the presence of right-censored failure time and missing covariates. Unlike the traditional Bayesian method that relies on an assessable likelihood function constructed by assigning specific distributions to the model’s random errors and missing covariates, the proposed BEL-based approach does not make any distributional assumption, yielding highly reliable and robust estimation results. In addition, by exploring the entire posterior distribution of unknowns, the BEL methods can construct credible intervals based on empirical percentiles and reveal the uncertainty of the Bayesian estimator in a straightforward manner, avoiding the challenge of deriving the asymptotic distribution of empirical likelihood ratio statistics. We develop an efficient Markov Chain Monte Carlo method coupled with the Metropolis-Hastings algorithm to conduct posterior inference and investigate the asymptotic behavior of the posterior distribution. Simulation studies show that the BEL procedure performs satisfactorily in various settings and consistently outperforms several existing methods. The application of the proposed method to a real-life dataset from the mouse leukemia study further confirms the practical utility of our method.

25CHI111: Statistics for Emerging Trends in Machine Learning

Advancing Fairness in Healthcare: A Universal Framework for Optimal Treatment Effect Estimation with Censored Data

Hongni Wang, ♦Junxi Zhang, Na Li, Linglong Kong, Bei Jiang, Xiaodong Yan

Shandong University of Finance and Economics, Concordia University, Shandong University of Finance and Economics, University of Alberta, University of Alberta, Xi'an Jiaotong University

In healthcare and precision medicine, estimating optimal treatment strategies for right-censored data while ensuring fairness across ethnic subgroups is crucial. This problem poses two primary challenges: measuring heterogeneous treatment effects (HTE) under various fairness constraints and addressing censoring mechanisms. In this talk, I will address these challenges by proposing a general framework for estimating HTE using nonparametric methods with user-controllable fairness constraints. The estimated HTE is then utilized to derive the optimal treatment strategy. Under mild regularization assumptions, the framework is theoretically grounded, exhibiting the double robustness property of the HTE estimator. Additionally, I will discuss how the derived optimal treatment strategy balances fairness and utility, shedding light on the well-known fairness-utility trade-off.

An adaptive model checking test for the functional linear model

Enze Shi, Yi Liu, Ke Sun, ♦Lingzhu Li, Linglong Kong

University of Alberta, University of Alberta, University of Alberta, Beijing University of Technology, University of Alberta

Numerous studies have been devoted to the estimation and inference problems for functional linear models (FLM). However, few works focus on model checking problem that ensures the reliability of results. Limited tests in this area do not have tractable null distributions or asymptotic analysis under alternatives. Also, the functional predictor is usually assumed to be fully observed, which is impractical. To address these problems, we propose an adaptive model checking test for FLM. It combines regular moment-based and conditional moment-based tests, and achieves model adaptivity via the dimension of a residual-based subspace. The advantages of our test are manifold. First, it has a tractable chi-squared null distribution and higher powers under the alternatives than its components. Second, asymptotic properties under different underlying models are developed, including the unvisited local alternatives. Third, the test statistic is constructed upon finite grid points, which incorporates the discrete nature of collected data. We develop the desirable relationship between sample size and number of grid points to maintain the asymptotic properties. Besides, we provide a data-driven approach to estimate the dimension leading to model adaptivity, which is promising in sufficient dimension reduction.

Some Recent Optimal Exact Confidence Intervals in Contingency Tables

♦weizhen wang

Beijing University of Technology

A general method, named the h-function method, has been introduced to obtain a $1-\alpha$ exact confidence interval. Using this method, any given confidence interval can be improved in the following ways: (i) an approximate interval, including a point estimator, can be modified to an exact interval; (ii) an exact interval can be refined to form a sub-interval of the previous one. Various applications of this method are discussed for estimating key parameters in contingency tables, including the difference between two independent proportions, relative risk, and odds ratio when two independent binomials are observed; the difference between two dependent proportions in a matched-pair experiment without or with missing values; and the risk difference and risk ratio in 2X2 contingency tables with a structural zero.

Online model averaging prediction

♦Jun Liao

Renmin University of China

For massive data, it is often difficult to describe the characteristics of the data by a single model and hence the associated prediction may not be desirable due to the model misspecification. In this paper, we develop the online model averaging predictions for massive data, which are particularly appropriate for the common scenario where the prediction model may be misspecified and the large-scale data are collected sequentially. The proposed methods do not need to use the whole data of the individual level and only need to utilize the current batch of data and some statistics based on the previous batches. Also, the new methods adapt to both the continuous response and discrete response cases, which include the common regression models, such as the ordinary linear regression, nonlinear regression, logistic regression and Poisson regression. The online model averaging estimators are shown to be asymptotically optimal. Further, the convergence rate of the weight estimator developed in terms of the squared prediction error is derived. The simulation study and real data analysis reveal that the proposed methods have the desirable finite sample performance.

25CHI024: Design and analysis of clinical studies

A Generalized Outcome-Adaptive Sequential Multiple Assignment Randomized Trial Design

Xue Yang, ♦Yu Cheng, Peter Thall, Wabdus Wahed

University of Pittsburgh, University of Pittsburgh, University of Texas MD Anderson Cancer Center, University of Rochester

A dynamic treatment regime (DTR) is a mathematical representation of a multistage decision process. When applied to sequential treatment selection in medical settings, DTRs are useful for identifying optimal therapies for chronic diseases such as AIDs, mental illnesses, substance abuse, and many cancers. Sequential multiple assignment randomized trials (SMARTs) provide a useful framework for constructing DTRs and providing unbiased between-DTR comparisons. A limitation of SMARTs is that they ignore data from past patients that may be useful for reducing the probability of exposing new patients to inferior treatments. In practice, this may result in decreased treatment adherence or dropouts. To address this problem, we propose a generalized outcome-adaptive (GO) SMART design that adaptively unbalances stage-specific randomization probabilities in favor of treatments observed to be more effective in previous patients. To correct for bias induced by outcome adaptive randomization, we propose G-estimators and inverse-probability-weighted estimators of DTR effects embedded in a GO-SMART and show analytically that they are consistent. We report simulation results showing that, compared to a SMART, Response-Adaptive SMART (Wang et al., 2022) and SMART with adaptive randomization (Cheung et al., 2015), a GO-SMART design treats significantly more patients with the optimal DTR and achieves a larger number of total responses while maintaining similar or better statistical power.

State-dependent sampling designs for prevalent cohort studies

♦*Leilei Zeng*

University of Waterloo

There is great interest in conducting studies to better understand patterns and trends of disease incidence, disease progression, and other comorbidities in populations. While birth cohorts, where individuals are followed prospectively starting at birth, yield high quality information about such transitions, they can be prohibitively expensive to conduct due to the long follow-up. As such, prevalent cohort studies are often preferred, although little attention has been given to the design of prevalent cohort studies for multistate life history processes. We propose a state-specific recruitment design with two different sampling schemes in the multistate framework. We consider the differing cost of recruiting individuals from different states and examine characteristic features of resulting minimum-cost designs.

Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

Xiangbin Xiangbin, ♦Wen Su, Zhisheng Ye, Xingqiu Zhao

The Hong Kong Polytechnic University, City University of Hong Kong, National University of Singapore, The Hong Kong Polytechnic University

Personalized treatment regimes tailored to account for individual characteristics have revolutionized the health care industry, offering substantial potential to minimize treatment risks and enhance patient survival. However, the current methods for estimating individualized treatment rules in multi-arm settings have limitations due to the lack of theoretical foundations and challenges related to model misspecification. To address these issues, we propose a novel generative learning approach called CG-Learning, that utilizes Wasserstein GAN to estimate the optimal decision rule that minimizes a risk measure for multi-armed treatment regimes. We derive important theoretical results of the proposed estimator including the nonasymptotic error bound for the estimated optimal value and an upper bound for the probability that the estimated decision rule is not the optimal treatment option. To evaluate performance of CG-Learning, we conduct extensive simulation studies under various scenarios, with comparisons to existing approaches. The proposed method is demonstrated using a dataset from the AIDS Clinical Trials Group

25CHI065: Recent Advances in Complex Data

Ordinary Differential Equation Models for a Collection of Discretized Functions

♦*Lingxuan Shao, Fang Yao*

Fudan University, Beijing University

The exploration of dynamic systems governed by Ordinary Differential Equations (ODEs) holds great interest in the field of statistics. Existing research mainly focuses on a single function. This study generalizes the scope to analyze a collection of functions observed at discretized times, with sampling frequencies varying from sparse to dense designs. The range of ODE models studied caters to diverse dynamic systems, and includes the complex non-linear and non-Lipschitz scenarios. We introduce a new concept named Functional Moment Method, a novel approach for parameter estimation within these ODE models and facilitating the recovery of curves for the discretely observed functions. Our numerical analysis underscores the methods applicability across various application fields, including sociology, physics, and epidemiology.

Regularized reduced-rank regression for structured output prediction

Heng Chen, Di-Rong Chen, ♦Kun Cheng, Yang Zhou

Capital University of Economics and Business, Beihang University, Beijing Jiaotong University, Beijing Normal University

Reduced-rank regression (RRR) has been widely used to strength the dependency among multiple outputs. In this talk, we introduce a regularized vector-valued RRR approach, which plays an important role in predicting multiple output variables with complex structures. The estimator of vector-valued RRR is obtained by minimizing the empirically squared reproducing kernel Hilbert space (RKHS) distances between output feature kernel and all r dimensional subspaces in vector-valued RKHS. The algorithm is implemented easily with kernel tricks. We establish the learning rate of vector-valued RRR estimator under mild assumptions. Moreover, as a reduced-dimensional approximation of output kernel regression function, the estimator converges to the output regression function in probability when the rank tends to infinity appropriately. It thus implies the consistency of structured predictor in general settings, especially in a misspecified case where the true regression function is not contained in the hypothesis space. Simulations and real data analysis are reported to illustrate the efficiency of the proposed method.

Two-Sample Distribution Tests in High Dimensions via Max-Sliced Wasserstein Distance and Bootstrapping

♦*Xiaoyu Hu, Zhenhua Lin*

Xi'an Jiaotong University, National University of Singapore

Two-sample hypothesis testing is a fundamental statistical problem for inference about two populations. In this paper, we construct a novel test statistic to detect high-dimensional distributional differences based on the max-sliced Wasserstein distance to mitigate the curse of dimensionality. By exploiting an intriguing link between the distance and suprema of empirical processes, we develop an effective bootstrapping procedure to approximate the null distribution of the test statistic. One distinctive feature of the proposed test is the ability to construct simultaneous confidence intervals for the max-sliced Wasserstein distances of projected distributions of interest. This enables not only the detection of global distributional differences but also the identification of significantly different marginal distributions between two populations without the need for additional tests. We establish the convergence of Gaussian and bootstrap approximations of the proposed test, based on which we show that the test is asymptotically valid and powerful as long as the considered max-sliced Wasserstein distance is adequately large. The merits of our approach are illustrated via simulated and real data examples.

Change-Points Detection and Support Recovery for Spatiotemporal Functional Data

♦*Decai Liang*

Nankai University

Large volumes of spatiotemporal data, including patterns of climatic variables, satellite images and FMRI data, usually exhibit inherent mean changes. Due to the complicated cross-covariance structure, the full covariance function is commonly described as a product of independent spatial covariance and temporal covariance, which is a mathematically convenient yet not always reflective assumption of the data. To remedy this, we propose a novel hypothesis test based on a more realistic assumption known as weak separability. We establish solid asymptotic theory to support this approach. Furthermore, we develop a comprehensive procedure for support recovery amidst the intricate correlations between space and time, effectively identifying true signals (locations with mean change) while controlling the false discovery rate. This represents the first work of support recovery within a spatiotemporal framework. Simulation studies and a Chinese precipitation data application validate the efficacy and enhanced power of our methodology on both change point detection and support recovery.

25CHI003: Advanced Statistical and Computational Methods for Microbiome and Metagenomics Data Analysis

Integrating functional and taxonomic profiles for microbiome biomarker identification and disease prediction

♦*Chan Wang, Huilin Li*

NYU School of Medicine, NYU School of Medicine

Recently, the microbiome has gained significant attention as a potential predictor of human diseases. However, identifying robust, validated, and powerful microbial biomarkers remains challenging due to the complexity of microbiome data, including both taxonomic and functional profiles. Studies have shown that taxonomic profiles typically offer greater predictive performance and are easier to apply in practical and clinical settings but exhibit higher variability. In contrast, functional profiles are more stable and interpretable in terms of biological mechanisms but tend to have lower predictive performance. In this study, we propose a robust microbial risk score (MRS) framework that integrates both taxonomic and functional profiles to identify a microbial sub-community capable of serving as biomarkers for disease susceptibility. Specifically, we first identify a sub-community of microbial taxa associated with disease using the taxonomic profile, following a similar approach to our MRS version 1. We then expand this sub-community by incorporating additional microbial taxa based on their functional similarities with the identified taxa and calculate the weighted diversities of the sub-community as the proposed MRSs. Through comprehensive real-data analyses using human microbiome datasets from the curatedMetagenomicData R package, we demonstrate the utility of the proposed MRS framework for disease prediction. Moreover, the incorporation of functional profiles can be seamlessly integrated into other predictive methods, such as random forests, to enhance predictive performance.

TEMPTED: time-informed dimensionality reduction for longitudinal microbiome studies

♦*Anru Zhang, Rungang Han, Yanan Zhao*

Duke University, Duke University, Duke University

Longitudinal studies are crucial for understanding complex microbiome dynamics and their link to health. In this talk, we introduce TEMPoral TEnsor Decomposition (TEMPTED), a time-informed dimensionality reduction method for high-dimensional longitudinal data that treats time as a continuous variable, effectively characterizing temporal information and handling varying temporal sampling. TEMPTED captures key microbial dynamics, facilitates beta-diversity analysis, and enhances reproducibility by transferring learned representations to new data. In simulations, it achieves 90% accuracy in phenotype classification, significantly outperforming existing methods. In real data, TEMPTED identifies vaginal microbial markers linked to term and preterm births, demonstrating robust performance across datasets and sequencing platforms.

Joint modeling of longitudinal and time-to-event outcomes under nested case-control sampling with application to TEDDY biomarker study

♦*Yanan Zhao, Jiyuan Hu*

Department of Population and Health, NYU Grossman School of Medicine, Department of Population and Health, NYU Grossman School of Medicine

The nested case-control (NCC) design provides a cost-effective alternative to full cohort biomarker studies while preserving statistical efficiency. Despite advances in joint modeling for longitudinal and time-to-event outcomes, few approaches address the unique challenges posed by NCC sampling, non-normally distributed biomarkers, and competing survival outcomes. Motivated by the TEDDY study, we propose "JM-NCC", a joint modeling framework designed for NCC studies with competing events. It integrates a generalized linear mixed-effects model for potentially non-normally distributed biomarkers with a cause-specific hazard model for competing risks. Two estimation methods are developed. fJM-NCC leverages NCC sub-cohort longitudinal biomarker data and full cohort survival and clinical metadata, while wJM-NCC uses only NCC sub-cohort data. Both simulation studies and an application to TEDDY microbiome dataset demonstrate the robustness and efficiency of the proposed methods.

25CHI010: Advances in Modern Statistical Methodologies: Robust Estimation, High-Dimensional Inference, and Innovative Biomedical Applications

Robust and Optimal Tensor Estimation via Robust Gradient Descent

♦*Xiaoyu Zhang*

Tongji University

Low-rank tensor models are widely used in statistics and machine learning. However, most existing methods rely heavily on the assumption that data follows a sub-Gaussian distribution. To address the challenges associated with heavy-tailed distributions encountered in real-world applications, we propose a novel robust estimation procedure based on truncated gradient descent for general low-rank tensor models. We establish the computational convergence of the proposed method and derive optimal statistical rates under heavy-tailed distributional settings of both covariates and noise for various low-rank models. Notably, the statistical error rates are governed by a local moment condition, which captures the distributional properties of tensor variables projected onto certain low-dimensional local regions. Furthermore, we present numerical results to demonstrate the effectiveness of our method.

Comparing MCP-MOD and Ordinal Linear Contrast Test in Dose Finding Clinical Trials: A Thorough Examination

Yaohua Zhang, ♦Ning Li, Naitee Ting

Boston University, Sanofi, StatsVita

The MCP-Mod approach (Pinheiro 2006) to confirm the proof of concept and identify the minimum effective dose (MED) in Phase II clinical trials was first introduced in the early 2000 era. The MCP-MOD method has significantly trans-formed the way dose finding studies are conducted. However, there are multiple issues with the dose(s) recommended by this method. Firstly, the selected doses are often not practical for manufacturing, which requires researchers to find the feasible dose along the significant dose response curve or return to the traditional pairwise test to find a manufacturable minimum effective dose. Secondly, the complexity of the MCP-MOD method makes it difficult to explain to a team. Thirdly, the algorithm encounters convergence issues. Last but not least, the per-formance of MCP-MOD is subject to whether the true dose-response relationship is close to one of the selected candidate models. In contrast, the Ordinal Linear Contrast Test (Zhang 2017) has demonstrated no such issues and is considered comparable to MCP-Mod in power and probability to select MED correctly to the MCP-MOD method. In this chapter, the authors conducted a thorough com-parison of the two methods through simulations and demonstrated the superiority of the Ordinal Linear Contrast Test.

Bayesian Design for Bridging Studies: Methods and Applications

♦Lichang Chen

Akeso Biopharma Inc.

A bridging study is an additional clinical trial that aims to extrapolate the efficacy and safety data of a drug from one population to another, typically from a foreign population to a local one. This approach is essential for accelerating the approval and availability of new drugs in different regions while minimizing the need for extensive and redundant clinical trials. Bayesian design methods have emerged as a powerful alternative for bridging studies. These methods incorporate historical data from a foreign population through the use of prior distributions, allowing for more efficient trial designs. In this talk, a short overview of power prior and robust mixture prior, will be provided. In addition, two exemplary studies illustrate the application and benefits of Bayesian design in bridging studies.

Bootstrap inference for high dimensional nonconvex penalised regression and post-selection least squares

♦Xiaoya XU, Stephen LEE

Shenzhen Polytechnic University, The University of Hong Kong

In the realm of high-dimensional linear regression, nonconvex penalised estimators have enjoyed increasing popularity due to their much acclaimed oracle property, which holds under assumptions weaker than those typically required for convex penalised estimators to enjoy the same property. However, validity of such oracle property of nonconvex penalisation and the accompanying inference tools is questionable in the presence of many weak signals and/or a few moderate signals, which may incur substantial biases. To address this issuewe aims at developing theoretically validcomputationally feasible, procedures for inference about the relationships between strong signals and response variables. In particular, the post-selection least squares method is found to improve on nonconvex penalised estimation, especially under heavy-tailed settings. Then, trustworthy bootstrap inference procedures based on nonconvex penalised estimators and their post-selection OLS estimators are developed to estimate their distribution under this flexible framework. The revised bootstrap method draws the bootstrap samples from the post-selection OLS model, and is shown to be valid under weaker conditions. Owing to its desirable theoretical properties, the residual bootstrap method based on post-selection least squares estimators is accurate generally, and can be as effective as normal approximation, even without assuming strong conditions on signal strength. Empirical results obtained from large-scale simulation and real data studies corroborate our theoretical findings.

25CHI018: Advancing Multi-platform and Multi-modal Omics Harmonization

Spotiphy enables single-cell spatial whole transcriptomics across an entire section

♦Jiyuan Yang

Department of Computational Biology, St. Jude Children's Research Hospital

Spatial transcriptomics (ST) has advanced our understanding of tissue regionalization by enabling the visualization of gene expression within whole-tissue sections, but current approaches remain plagued by the challenge of achieving single-cell resolution without sacrificing whole-genome coverage. Here we present Spotiphy (spot imager with pseudo-single-cell-resolution histology), a computational toolkit that transforms sequencing-based ST data into single-cell-resolved whole-transcriptome images. Spotiphy delivers the most precise cellular proportions in extensive benchmarking evaluations. Spotiphy-derived inferred single-cell profiles reveal astrocyte and disease-associated microglia regional specifications in Alzheimer’s disease and healthy mouse brains. Spotiphy identifies multiple spatial domains and alterations in tumor–tumor microenvironment interactions in human breast ST data. Spotiphy bridges the information gap and enables visualization of cell localization and transcriptomic profiles throughout entire sections, offering highly informative outputs and an innovative spatial analysis pipeline for exploring complex biological systems.

MODE: high-resolution digital dissociation with deep multimodal autoencoder

Jiao Sun, Tong Lin, Kyle Smith, Wei Zhang, Paul Northcott, ♦Qian Li

St. Jude Children's Research Hospital, St. Jude Children's Research Hospital, St. Jude Children's Research Hospital, University of Central Florida, St. Jude Children's Research Hospital, St. Jude Children's Research Hospital

Single-cell technologies enable high-resolution profiling of molecular dynamics in developmental and cancer biology. But heterogeneity and complexity of tumors may hinder the lineage cell mapping in developmental origins or dissection of tumor microenvironment, requiring digital dissociation of bulk tissues. Many deconvolution methods focus on transcriptomic assay using scRNA-seq as reference, not easily applicable to other omics due to ambiguous cell markers and unexpected biological difference between reference and target tissues. Here, we present MODE, a multimodal autoencoder pipeline linking multi-dimensional molecular features to jointly predict personalized multi-omic profiles and estimate modality-specific cellular compositions, using pseudo-bulk data constructed by internal non-transcriptomic signature matrix recovered from target tissues and external scRNA-seq reference. The accuracy of MODE was evaluated through extensive simulation experiments generating realistic multi-omic data from distinct tissue types. MODE outperformed seven deconvolution pipelines with superior generalizability and enhanced fidelity across five independent datasets, elucidating multi-omic signatures for disease developmental origins, evolution, subtyping, and prognosis.

Inferring cell type-specific co-methylation networks from single-cell DNA methylation data

♦Jiebiao Wang

University of Pittsburgh

Single-cell DNA methylation (scDNAm) technologies enable detailed studies of cell type-specific epigenetic regulation, yet co-methylation network inference remains challenging due to data sparsity, zero-one inflation, and complex dependencies. We present a novel statistical framework for cell type-specific co-methylation inference using a zero–one-inflated beta copula model. Our approach flexibly captures marginal and joint methylation distributions while accounting for sparsity and leveraging copulas for dependence modeling. By integrating cell type information, we identify distinct epigenetic modules and regulatory pathways. We demonstrate the utility of our method on simulated and real scDNAm data, uncovering biologically meaningful co-methylation patterns linked to cell function. This work offers a powerful tool to decipher the epigenetic landscape at single-cell resolution and illuminates cell type-specific regulatory mechanisms.

A deconvolution framework that uses single-cell sequencing plus a small benchmark data set for accurate analysis of cell type ratios in complex tissue samples

Shuai Guo, ♦Xiaoqian Liu, Xuesen Cheng, Rui Chen, Wenyi Wang

The University of Texas MD Anderson Cancer Center, University of California, Riverside, Baylor College of Medicine, Baylor College of Medicine, The University of Texas MD Anderson Cancer Center,

Bulk deconvolution with single-cell/nucleus RNA-seq data is critical for understanding heterogeneity in complex biological samples, yet the technological discrepancy across sequencing platforms limits deconvolution accuracy. To address this, we utilize an experimental design to match inter-platform biological signals, hence revealing the technological discrepancy, and then develop a deconvolution framework called DeMixSC using this well-matched, that is, benchmark, data. Built upon a novel weighted nonnegative least-squares framework, DeMixSC identifies and adjusts genes with high technological discrepancy and aligns the benchmark data with large patient cohorts of matched-tissue-type for large-scale deconvolution. Our results using two benchmark data sets of healthy retinas and ovarian cancer tissues suggest much-improved deconvolution accuracy. Leveraging tissue-specific benchmark data sets, we applied DeMixSC to a large cohort of 453 age-related macular degeneration patients and a cohort of 30 ovarian cancer patients with various responses to neoadjuvant chemotherapy. Only DeMixSC successfully unveiled biologically meaningful differences across patient groups, demonstrating its broad applicability in diverse real-world clinical scenarios. Our findings reveal the impact of technological discrepancy on deconvolution performance and underscore the importance of a well-matched data set to resolve this challenge. The developed DeMixSC framework is generally applicable for accurately deconvolving large cohorts of disease tissues, including cancers, when a well-matched benchmark data set is available.

25CHI019: Advancing Multi-Regional Clinical Trials: Methodology and Application Considerations for Ensuring Global Representation, Regulatory Harmonization, and Ethical Integrity

Consistency Assessment of Treatment Effect in Multi-Regional Clinical Trials in the Presence of Treatment Effect Heterogeneity

♦Menggang Yu, Kunhai Qing

University of Michigan, East China Normal University

Multi-regional clinical trial (MRCT) has been common practice for drug development and global registration. Regional consistency assessment methods have not adequately incorporated possible treatment effect heterogeneity. In this paper, we begin by revisiting the concept of consistency through the lens of the conditional average treatment effect function. Next, we propose an extended framework for defining consistency criteria in presence of covariate shift that may commonly exist between different regions. We develop a principled method for consistency assessment that can be prespecified at the protocol development stage. Results from extensive numerical simulations using various types of outcomes demonstrate the effectiveness of the proposed method in identifying consistent scenarios.

Considerations of China joining MRCTs - A case study in Central Nervous System (CNS) Therapeutic Area

♦Heli Gao, Hui Wang

Boehringer Ingelheim (China) Investment Co., Ltd., Boehringer Ingelheim (China) Investment Co., Ltd.

The integration of China into Multi-Regional Clinical Trials (MRCTs) represents a significant shift in the global landscape of drug development and regulatory approval. MRCTs are pivotal in assessing the efficacy and safety of new therapeutic interventions across diverse populations and geographies, thereby expediting the availability of innovative treatments worldwide. As China emerges as a major player in the pharmaceutical sector, its participation in MRCTs offers a unique set of considerations that must be addressed to harness the full potential of this collaboration. This presentation will use a case in Central Nervous System (CNS) Therapeutic Area to provide a comprehensive overview of the strategic considerations that stakeholders must navigate to realize the benefits while mitigating the risks associated with this global endeavor.

Regional consistency evaluation and sample size calculation under two MRCTs

Kunhai Qing, Xinru Ren, Shuping Jiang, Ping Yang, Menggang Yu, ♦Jin Xu

East China Normal University, East China Normal University, MSD China, MSD China, University of Michigan School of Public Health, East China Normal University

Multi-regional clinical trial (MRCT) has been common practice for drug development and global registration. The FDA guidance 'Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products Guidance for Industry' \citep{FDASEE} requires that substantial evidence of effectiveness of a drug/biologic product to be demonstrated for market approval. In the situations where two pivotal MRCTs are needed to establish effectiveness of a specific indication for a drug or biological product, a systematic approach of consistency evaluation for regional effect is crucial. In this paper, we first present some existing regional consistency evaluations in a unified way that facilitates regional sample size calculation under the simple fixed effect model. Second, we extend the two commonly used consistency assessment criteria of \citet{japan_criteria} in the context of two MRCTs and provide their evaluation and regional sample size calculation. Numerical studies demonstrate the proposed regional sample size attains the desired probability of showing regional consistency. A hypothetical example is presented to illustrate the application. We provide an R package for implementation.

Practical Considerations on Bayesian Hierarchical Models to Support Regional Effect Evaluation in Multi-Regional Clinical Trials

♦*Renxin Lin*

Novartis China

Multi-regional clinical trial (MRCT) data is pivotal for the simultaneous evaluation and approval of new drugs across different regions. Evaluating regional effects has been a challenging task. Several guidelines have been released on MRCTs, including the recently released draft guidance from the China Center for Drug Evaluation (CDE) on Benefit-Risk Assessment. This guidance discusses key considerations in the design and implementation of MRCTs, including the evaluation of regional differences and consistency assessment across regions. The Bayesian approach to estimate regional differences is also mentioned in the guidance and has been widely discussed in the literature.This presentation will provide an overview of Bayesian Hierarchical Models (BHM), highlighting their ability to offer a structured approach to linking treatment effects across various studies and subgroups. This methodology provides a shrinkage estimate of regional effects, which is particularly beneficial in addressing the challenges posed by heterogeneous treatment effects (HTE) observed in different subgroups. By leveraging shrinkage estimation, BHM provides more precise and reliable estimates of treatment effects. Additionally, the presentation will discuss practical considerations for implementing BHM in MRCTs. Lastly, the use of BHM will be discussed within the framework of the totality of evidence, ensuring a comprehensive and robust evaluation of regional effects.

25CHI025: Dimension Reduction Methods

Fast fitting of Gaussian mixture model via dimension reduction

♦*Yin Jin, Wei Luo*

Zhejiang University, Zhejiang University

The Gaussian Mixture Model (GMM) stands out as a widely applied clustering framework. Commonly, the maximal likelihood approach to fit GMM requires solving a non-convex optimization, which is computationally challenging especially for large-dimensional data. To address the problem, we propose a two-step approach to utilize the intrinsic low-dimensional structure in GMM under additional constraints on the heterogeneity of GMM. In the first step, we use a simple method to recover the low-dimensional data, given which the rest of data are normally distributed and thus redundant for clustering. We then fit GMM using the reduced data in the second step, which is computationally more feasible than the original GMM due to the lower dimensionality. Under the sparsity assumption on the clustering pattern, our approach can be generalized under the ultrahigh-dimensional settings. It can also be embedded under a general framework of sufficient dimension reduction, which encompasses more methods to recover the low-dimensional structure of GMM in the future. The numerical studies show that our algorithm significantly accelerates the computation compared to the existing methods.

Robust Sliced Inverse Regression: Optimal Estimation for Heavy-Tailed Data in High Dimensions

♦*Jing Zeng, Keqian Min, Qing Mai*

University of Science and Technology of China, IBM, Florida State University

Sliced inverse regression (SIR) is a flexible modeling tool that effectively reduces dimensions to reveal the complicated mechanism behind data. In recent years, SIR has been generalized to high dimensions in a variety of ways. However, all existing methods rely on the light-tailed assumption for predictors, which is frequently violated in real life. To tackle ubiquitous heavy-tailed data, we propose a novel robust SIR method, referred to as ROSE, that scales well with high dimensions and heavy tails simultaneously. We start with an adaptive distribution model that explicitly incorporates heavy tails and covers many popular distributions as special cases. Then ROSE leverages a new elegant invariance result to convert the original SIR problem to a less challenging one on a set of latent light-tailed predictors. We rigorously show that ROSE admits the same minimax optimal convergence rate as existing light-tailed methods even when we only have finite second moments. ROSE is also computationally efficient compared to existing robust methods in that no extra tuning parameter selection is required to overcome the heavy-tailedness. Extensive empirical studies are conducted to support the theoretical results.

A Reduced-Rank Factor Model for Panel Data

Mingke Zhang, ♦*Yingcun Xia*

National University of Singapore, National University of Singapore

We consider a new framework, the Reduced-Rank Factor Model for Panel Data, to address the dual challenges of individual heterogeneity and unobserved cross-sectional dependence in panel data analysis. We address estimation bias arising from the presence of heterogeneous coefficients and interactive fixed effects, and propose a consistent selection criterion for jointly determining the number of reduced-rank components and latent factors.

25CHI028: High dimensional statistics inference

Nonlinear Principal Component Analysis with Random Bernoulli Features for Process Monitoring

ke Chen, ♦Dandan Jiang

Xi'an Jiaotong University, Xi'an Jiaotong University

The process generates substantial amounts of data with highly complex struc- tures, leading to the development of numerous nonlinear statistical methods. However, most of these methods rely on computations involving large-scale dense kernel matrices. This dependence poses significant challenges in meet- ing the high computational demands and real-time responsiveness required by online monitoring systems. To alleviate the computational burden of dense large-scale matrix multiplication, we incorporate the bootstrap sam- pling concept into random feature mapping and propose a novel random Bernoulli principal component analysis method to efficiently capture nonlin- ear patterns in the process. We derive a convergence bound for the kernel matrix approximation constructed using random Bernoulli features, ensuring theoretical robustness. Subsequently, we design four fast process monitoring methods based on random Bernoulli principal component analysis to ex- tend its nonlinear capabilities for handling diverse fault scenarios. Finally, numerical experiments and real-world data analyses are conducted to eval- uate the performance of the proposed methods. Results demonstrate that the proposed methods offer excellent scalability and reduced computational complexity, achieving substantial cost savings with minimal performance loss compared to traditional kernel-based approaches.

Portmanteau statistics for high-dimensional vector moving average processes

♦Chi Yao, Zeqin Lin, Xuejun Wang, Yiming Liu, Guangming Pan

Nanyang Technological University, Nanyang Technological University, Anhui University, Jinan University, Nanyang Technological University,

Consider high-dimensional vector moving average processes of order q (VMA(q)) in the high-dimensional regime where both the dimension p and the sample size n diverge to infinity. We establish multivariate central limit theorems for the traces of the differences between the symmetrized sample autocovariance matrices at multiple time lags. Building on these results, we propose a novel portmanteau statistic in the spirit of the Ljung-Box statistic. This statistic, along with its variant at a single time lag, can be applied to test the order of VMA(q) processes and, in particular, to test whether the process is a vector white noise. Our proposed methods are valid under the growth assumption $cn \leq p \ll n^{4/3}$ where $c > 0$ is some (small) constant, and notably, they circumvent the need to estimate the coefficient matrices or the fourth moments of the innovation terms. Numerical simulation results under both large-sample and small-sample scenarios confirm the accuracy and practical relevance of the proposed procedures.

Inference in Randomized Least Squares and PCA via Normality of Quadratic Forms

Leda Wang, ♦Zhixiang Zhang, Edgar Dobriban

Yale University, University of Macau, University of Pennsylvania

Randomized algorithms can be used to speed up the analysis of large datasets. In this paper, we develop a unified methodology for statistical inference via randomized sketching or projections in two of the most fundamental problems in multivariate statistical analysis: least squares and PCA. The methodology applies to fixed datasets-i.e., is data-conditional-and the only randomness is due to the randomized algorithm. We propose statistical inference methods for a broad range of sketching distributions, such as the subsampled randomized Hadamard transform (SRHT), Sparse Sign Embeddings (SSE) and CountSketch, sketching matrices with i.i.d. entries, and uniform subsampling. To our knowledge, no comparable methods are available for SSE and for SRHT in PCA. Our novel theoretical approach rests on showing the asymptotic normality of certain quadratic forms. As a contribution of broader interest, we show central limit theorems for quadratic forms of the SRHT, relying on a novel proof via a dyadic expansion that leverages the recursive structure of the Hadamard transform. Numerical experiments using both synthetic and empirical datasets support the efficacy of our methods, and in particular suggest that sketching methods can have better computation-estimation tradeoffs than recently proposed optimal subsampling methods.

Distribution-Free and Model-Agnostic Changepoint Detection with Finite-Sample Guarantees

Xiaolong Cui, Haoyu Geng, ♦Guanghui Wang, Zhaojun Wang, Changliang Zou

Nankai University, Nankai University, Nankai University, Nankai University, Nankai University,

We introduce ART, a distribution-free and model-agnostic framework for changepoint detection that provides finite-sample guarantees. ART transforms independent observations into real-valued scores via a symmetric function, ensuring exchangeability in the absence of changepoints. These scores are then ranked and aggregated to detect distributional changes. The resulting test offers exact Type-I error control, agnostic to specific distributional or model assumptions. Moreover, ART seamlessly extends to multi-scale settings, enabling robust multiple changepoint estimation and post-detection inference with finite-sample error rate control. By locally ranking the scores and performing aggregations across multiple prespecified intervals, ART identifies changepoint intervals and refines subsequent inference while maintaining its distribution-free and model-agnostic nature. This adaptability makes ART as a reliable and versatile tool for modern changepoint analysis, particularly in high-dimensional data contexts and applications leveraging machine learning methods.

25CHI029: High dimensional statistics inference (II)

Huber Principal Component Analysis for large-dimensional factor models

Yong He, Lingxiao Li, ♦Dong Liu, Wen-Xin Zhou

Institute for Financial Studies, Shandong University, Department of Data Science and Artificial Intelligence, Hong Kong Polytechnic University, School of Physical and Mathematical Sciences, Nanyang Technological University, College of Business Administration, University of Illinois at Chicago

Factor models have been widely used in economics and finance. However, the heavy-tailed nature of macroeconomic and financial data is often neglected in statistical analysis. To address this issue, we propose a robust approach to estimate factor loadings and scores by minimizing the Huber loss function, which is motivated by the equivalence between conventional

Principal Component Analysis (PCA) and the constrained least squares method in the factor model. We provide two algorithms that use different penalty forms. The first algorithm involves an element-wise-type Huber loss minimization, solved by an iterative Huber regression algorithm. The second algorithm, which we refer to as Huber PCA, minimizes the $\ell_2$-norm-type Huber loss and performs PCA on the weighted sample covariance matrix. We examine the theoretical minimizer of the element-wise Huber loss function and demonstrate that it has the same convergence rate as conventional PCA when the idiosyncratic errors have bounded second moments. We also derive their asymptotic distributions under mild conditions. Moreover, we suggest a consistent model selection criterion that relies on rank minimization to estimate the number of factors robustly. We showcase the benefits of the proposed two algorithms through extensive numerical experiments and a real macroeconomic data example. An R package named "HDRFA" 1 has been developed to conduct the proposed robust factor analysis.

Revisiting the Spanning Puzzle of Bond Returns: A Unified Econometric Inference based on Predictive Quantile Regression

♦*Xiaohui Liu, Xinyi Wei*, Wei Long

Jiangxi University of Finance and Economics, Jiangxi University of Finance and Economics, Department of Economics, Tulane University

The validity of the "spanning hypothesis" remains a prominent topic of discussion within the bond market literature. In this paper, we investigate the spanning hypothesis within the framework of predictive quantile regression, a perspective that has received limited attention in the existing research. We propose a unified inferential strategy that accommodates predictors with varying levels of persistence. Additionally, we introduce a random weighted bootstrap procedure to circumvent the need for estimating the conditional density function, thereby enhancing the robustness of our method. Numerical simulations demonstrate that the proposed approach delivers strong finite-sample performance across a wide range of scenarios. Lastly, we apply our method to revisit several previously published studies, thereby enriching the empirical insights derived from predictive mean regression.

Identify the source of spikes: factor or mixture?

♦*Yiming Liu*

Jinan University

We consider the problem of identifying the pattern of latent variables in high-dimensional linear latent variable models, which can also be interpreted as determining the source of spiked singular values in the data matrix. Specifically, we test whether the latent variables are continuous or categorical, a distinction which is crucial for data interpretation but challenging when the dimensionality is comparable to the sample size. To address this inference problem, we analyze the asymptotic behavior of empirical measures associated with singular vectors corresponding to large spiked singular values. Leveraging these insights, we propose a novel test statistic based on the eigenvector quantile differences and establish its theoretical performance under the null hypothesis. Simulation studies and real data analyses for breast cancer and glioblastoma gene expression datasets demonstrate the effectiveness and practical utility of our method.

Identifying the structure of high-dimensional time series via eigen-analysis

♦*Bo Zhang, Jiti Gao, Guangming Pan, Yanrong Yang*

University of Science and Technology of China, Monash University, Nanyang Technological University, Australian National University

Cross-sectional structures and temporal tendency are important features of high dimensional time series. Based on eigen-analysis on sample covariance matrices, we propose a novel approach to identifying four popular structures of high-dimensional time series, which are grouped in terms of factor structures and stationarity. The proposed three-step method includes:

(1) a ratio statistic of empirical eigenvalues;

(2) a projected Augmented Dickey-Fuller Test;

(3) a new unit-root test based on the largest empirical eigenvalues.

We develop asymptotic properties for these three statistics to ensure the feasibility of the whole identifying procedure. Finite sample performances are illustrated via various simulations. We also analyze U.S. mortality data, U.S. house prices and income, and U.S. sectoral employment, all of which possess cross–sectional dependence and non-stationary temporal dependence. It is worth mentioning that we also contribute to statistical justification for the benchmark paper by Lee and Carter [32] in mortality forecasting.

25CHI031: Innovations in Causal Inference and Statistical Methods for Complex Data Structures

The synthetic instrument: From sparse association to sparse causation

♦*Dingke Tang, Dehan Kong, Linbo Wang*

Washington University in St. Louis, University of Toronto, University of Toronto

In many observational studies, researchers are often interested in studying the effects of multiple exposures on a single outcome. Standard approaches for high-dimensional data such as the lasso assume the associations between the exposures and the outcome are sparse. These methods, however, do not estimate the causal effects in the presence of unmeasured confounding. In this paper, we consider an alternative approach that assumes the causal effects in view are sparse. We show that with sparse causation, the causal effects are identifiable even with unmeasured confounding. At the core of our proposal is a novel device, called the synthetic instrument, that in contrast to standard instrumental variables, can be constructed using the observed exposures directly. We show that under linear structural equation models, the problem of causal effect estimation can be formulated as an l0-penalization problem, and hence can be solved efficiently using off-the-shelf software. Simulations show that our approach outperforms state-of-art methods in both low-dimensional and high-dimensional settings. We further illustrate our method using a mouse obesity dataset.

Geodesic Causal Inference

Daisuke Kurisu, ♦*Yidong Zhou, Taisuke Otsu, Hans-Georg Müller*

Center for Spatial Information Science, The University of Tokyo, Department of Statistics, University of California, Davis, Department of Economics, London School of Economics, Department of Statistics, University of California, Davis

Adjusting for confounding and imbalance when establishing statistical relationships is an increasingly important task, and causal inference methods have emerged as the most popular tool to achieve this. Causal inference has been developed mainly for regression relationships with scalar responses and also for distributional responses. We introduce here a general framework for causal inference when responses reside in general geodesic metric spaces, where we draw on a novel geodesic calculus that facilitates scalar multiplication for geodesics and the quantification of treatment effects through the concept of geodesic average treatment effect. Using ideas from Fréchet regression, we obtain a doubly robust estimation of the geodesic average treatment effect and results on consistency and rates of convergence for the proposed estimators. We also study uncertainty quantification and inference for the treatment effect. Examples and practical implementations include simulations and data illustrations for responses corresponding to compositional responses as encountered for U.S. statewise energy source data, where we study the effect of coal mining, network data corresponding to New York taxi trips, where the effect of the COVID-19 pandemic is of interest, and the studying the effect of Alzheimer's disease on connectivity networks.

Generalized Independence Test for Modern Data

♦*Mingshuo Liu, Doudou Zhou, Hao Chen*

UCDavis, NUS, UCDavis

The test of independence is widely needed in various applications of modern data analysis. However, traditional methods often struggle with the complex dependency structures found in high-dimensional data. To address this challenge, we propose a novel test statistic that captures intricate relationships using similarity and dissimilarity information from the data. The statistic exhibits strong power across a broad range of alternatives for high-dimensional data, as demonstrated in extensive simulation studies. Under mild conditions, we show that the new test statistic converges to the χ^2_4 distribution under the permutation null distribution, ensuring straightforward type I error control. Furthermore, our

research advances the moment method to prove the joint asymptotic normality of multiple double-indexed permutation statistics. We showcase the practical utility of this new test with an application to the Genotype-Tissue Expression dataset, where it effectively identifies associations between human tissues.

Modeling Interval-Censored Outcome Data with a Potentially Interval-Censored Covariate

♦*Dongdong Li, Yue Song, Wenbin Lu, Huldrych Gunthard, Roger Kouyos, Rui Wang*

Harvard Medical School

Identifying predictors for viral rebound trajectories after antiretroviral therapy (ART) interruption is central to HIV cure research. Motivated by the need to assess whether the time to achieve viral suppression after ART initiation predicts the time to viral rebound following ART interruption, we investigate modeling approaches that relate an interval-censored outcome (e.g., time to viral rebound) and an interval-censored covariate (e.g., time to viral suppression). We develop estimation and inference procedures for fitting a proportional hazards regression model when both the outcome and a covariate are interval-censored, using an Expectation-Maximization algorithm without making parametric distributional assumptions about baseline hazard functions. Given that some participants experienced multiple episodes of ART

initiation and interruption, we extend the proposed method to account for the clustering effect of multiple observations from the same participant. We evaluate the finite-sample performance of the proposed method for both independent and clustered data settings through simulation studies. To illustrate, we assess the association between time to viral suppression after ART initiation and time to viral rebound after ART interruption using data from the Zurich Primary HIV Infection cohort.

25CHI034: Innovations in Nonparametric and Functional Data Analysis

Semiparametric mixture regression for asynchronous longitudinal data using multivariate functional principal component analysis

♦*Yehua Li, Ruihan Lu, Weixin Yao*

University of California, Riverside, US FDA, University of California, Riverside

The transitional phase of menopause induces significant hormonal fluctuations, exerting a profound influence on the long-term well-being of women. In an extensive longitudinal investigation of women’s health during mid- life and beyond, known as the Study of Women’s Health Across the Nation (SWAN), hormonal biomarkers are repeatedly assessed, following an asynchronous schedule compared to other error-prone covariates, such as physical and cardiovascular measurements. We conduct a subgroup analysis of the SWAN data employing a semiparametric mixture regression model, which allows us to explore how the relationship between hormonal responses and other time-varying or time-invariant covariates varies across subgroups. To address the challenges posed by asynchronous scheduling and measurement errors, we model the time-varying covariate trajectories as functional data with reduced-rank Karhunen-Lo eve expansions, where splines are employed to capture the mean and eigenfunctions. Treating the latent subgroup membership and the functional principal component (FPC) scores as missing data, we propose an Expectation-Maximization (EM) algorithm to effectively fit the joint model, combining the mixture regression for the hormonal response and the FPC model for the asynchronous, time-varying covariates. In addition, we explore data-driven methods to determine the optimal number of subgroups within the population. Through our comprehensive analysis of the SWAN data, we unveil a crucial subgroup structure within the aging female population, shedding light on important distinctions and patterns among women undergoing menopause.

Estimation and Inference for Nonparametric Expected Shortfall Regression over RKHS

Myeonghun Yu, ♦*Yue Wang, Siyu Xie, Kean Ming Tan, Wenxin Zhou*

University of Michigan, University of Science and Technology of China, Northwestern University, University of Michigan, University of Illinois at Chicago,

Expected shortfall (ES) has emerged as an important metric for characterizing the tail behavior of a random outcome, specifically associated with rarer events that entail severe consequences. In climate science, the threats of flooding and heatwaves loom large, impacting natural environments and human communities. In actuarial studies, a key observation in modeling insurance claim sizes is that features exhibit distinct effects in explaining small and large claims. This article concerns nonparametric expected shortfall regression as a class of statistical methods for tail learning. These methods directly target upper/lower tail averages and will empower practitioners to address complex questions that are beyond the reach of mean regression based approaches. Using kernel ridge regression, we introduce a two-step nonparametric ES estimator that involves a plugged-in quantile function estimate without sample-splitting. We provide non-asymptotic estimation and Gaussian approximation error bounds, depending explicitly on the effective dimension, sample size, regularization parameters, and quantile estimation error. To construct pointwise confidence bands, we propose a fast multiplier bootstrap procedure and establish its validity. We demonstrate the finite-sample performance of the proposed methods through numerical experiments and an empirical study aimed at examining the heterogeneous effects of different air pollutants and meteorological factors on average and high PM2.5 concentration. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Variable Selection and Minimax Prediction in High-dimensional Functional Linear Models

♦Xingche Guo, Yehua Li, Tailen Hsing

University of Connecticut, University of California, Riverside, University of Michigan

High-dimensional functional data have become increasingly prevalent in modern applications such as high-frequency financial data and neuroimaging data analysis. We investigate a class of high-dimensional linear regression models, where each predictor is a random element in an infinite-dimensional function space, and the number of functional predictors p can potentially be ultra-high. Assuming that each of the unknown coefficient functions belongs to some reproducing kernel Hilbert space (RKHS), we regularize the fitting of the model by imposing a group elastic-net type of penalty on the RKHS norms of the coefficient functions. We show that our loss function is Gateaux sub-differentiable, and our functional elastic-net estimator exists uniquely in the product RKHS. Under suitable sparsity assumptions and a functional version of the irrepresentable condition, we derive a non-asymptotic tail bound for variable selection consistency of our method. Allowing the number of true functional predictors q to diverge with the sample size, we also show a post-selection refined estimator can achieve the oracle minimax optimal prediction rate. The proposed methods are illustrated through simulation studies and a real-data application from the Human Connectome Project.

Bias-Correction and Test for Mark-Point Dependence with Replicated Marked Point Processes

Ganggang Xu, Jingfei Zhang, Yehua Li, ♦Yongtao Guan

University of Miami, Emory University, UC Riverside, Chinese University of Hong Kong, Shenzhen

Mark-point dependence plays a critical role in research problems that can be fitted into the general framework of marked point processes. In this work, we focus on adjusting for mark-point dependence when estimating the mean and covariance functions of the mark process, given independent replicates of the marked point process. We assume that the mark process is a Gaussian process and the point process is a log-Gaussian Cox process, where the mark-point dependence is generated through the dependence between two latent Gaussian processes. Under this framework, naive local linear estimators ignoring the mark-point dependence can be severely biased. We show that this bias can be corrected using a local linear estimator of the cross-covariance function and establish uniform convergence rates of the bias-corrected estimators. Furthermore, we propose a test statistic based on local linear estimators for mark-point independence, which is shown to converge to an asymptotic normal distribution in a parametric root n−convergence rate. Model diagnostics tools are developed for key model assumptions and a robust functional permutation test is proposed for a more general class of mark-point processes. The effectiveness of the proposed methods is demonstrated using extensive simulations and applications to two real data examples.

25CHI043: Lifetime Data Analysis

Semiparametric analysis of recurrent disease status data collected at intermittent follow-up

Yong-Chen Huang, ♦Shu-Hui Chang

National Taiwan University, National Taiwan University

In clinical studies of chronic diseases, the primary endpoint (e.g, recurrent event) is collected intermittently at clinical visits in electronic medical records, rather than continuously over time, so the exact time of disease recurrence may not be observed. The timing and frequency of clinical visits may depend on the patient's past disease history and relevant clinical features. Semiparametric gap time hazards models for recurrent events are considered for evaluating the covariate effects. The estimation method for covariate effects is developed using the information on serial recurrent event status at clinical visit process, in which the weighted approach is adopted to address the informative clinical visits. The asymptotic normality of the proposed estimators is investigated. The finite-sample performance of the proposed method is assessed by simulation studies. The proposed method is further illustrated by a clinical example for patients with thyroid cancer.

A Varying-coefficient Additive Hazard Model for Recurrent Events Data

♦Jialiang Li

NUS

The additive hazard model, which focuses on risk differences rather than risk ratios, has been widely applied in practice. In this paper, we consider an additive hazard model with varying coefficients to analyze recurrent events data. The model allows for both varying and constant coefficients. We first propose an estimating equation-based approach with spline basis smoothing for all functional coefficients. Then, we provide theoretical justifications for the resulting estimates, including consistency, rate of convergence and asymptotic distribution. Furthermore, we construct a Cram\'{e}r-von Mises test procedure to investigate whether the functional coefficients should be treated as constant, and its asymptotic null distribution is also derived. Extensive simulation experiments are conducted to evaluate the finite-sample performance of the proposed approaches. A Chronic Granulotamous Disease data set was analyzed to illustrate our methodology.

Deep Nonparametric Inference for Censored Data

Wen Su, Qiang Wu, Kin-Yat Liu, Guosheng Yin, Jian Huang, ♦Xingqiu Zhao

City University of Hong Kong, The Hong Kong Polytechnic University, The Chinese University of Hong Kong, The University of Hong Kong, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

In this article, we propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the convergence rate and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Notably, we design a new test specifically tailored for testing nonparametric Cox models. The consistency of these tests is derived. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

Time-Adapted Exponential Models for Recurrent Covariates in Survival Analysis

Guangyu Yang, Boxian Wei, ♦Min Zhang

Institute of Statistics and Big Data, Renmin University of China, Vanke School of Public Health, Tsinghua University, Vanke School of Public Health, Tsinghua University

Recurrent events — such as hospitalizations, adverse events, and asthma attacks —are common in clinical studies. While statistical methods for analyzing recurrent events as outcomes are well established, limited research exists on modeling recurrent events as exposures for time-to-event outcomes (e.g., mortality). Our study is motivated by the need to assess the effect of bleeding events on mortality in patients who have received a left ventricular assist device (LVAD). Bleeding is a major post-LVAD complication and often occurs multiple times. Current statistical approaches have two key limitations: (1) they typically consider only the occurrence of bleeding, ignoring subsequent events, and (2) they often rely on the unrealistic assumption of a constant, permanent effect. To address these limitationsin this talk, I will introduce a novel time-adaptive exponential model to evaluate the impact of recurrent, time-varying bleeding events on post-LVAD mortality.

25CHI045: Model fairness and challenges and development of statistical models

Fairness-Constrained Optimal Model Averaging with High-Dimensional Sparsity Learning

Zeyu Chen, ♦Wei Qian, Bintong Chen

University of Delaware, University of Delaware, University of Delaware

Model fairness issues have received considerable attention recently in machine learning community. The popular class of generalized linear models such as logistic regression have seen various developments to improve fairness. Despite the progress, studies remain very limited in high-dimensional settings without adequate understanding from statistical perspectives. In this work, we propose a novel fairness-constrained model averaging algorithm for generalized linear models that can aggregate a large number of sparse model candidates to generate asymptotically fair modeling solutions. It is shown to achieve near optimal estimation risk in combining for estimation improvement, including flexible scenarios that a true model is not among the feasible candidates or is not necessarily representable by a single sparse model. To facilitate practical applications for the model averaging approach, we further propose a new fairness-assisted stepwise sparsity learning method to help generate potentially fair model candidates. In addition, the fairness-assisted stepwise method with model selection maintains consistency properties when the true model is among the sparse feasible candidates, showing delicate distinction of combining for estimation improvement versus adaptation.

Leveraging Multiple Endpoints to Estimate and Identify Subgroup-Specific Treatment Effects

♦Tom Chen, Emma Smith, Nick Birk, Rui Wang

Harvard Medical School, Harvard T. H. Chan School of Public Health, Harvard T. H. Chan School of Public Health, Harvard T. H. Chan School of Public Health & Harvard Medical School

Randomized controlled trials (RCTs) commonly report average treatment effects (ATEs), which may mask variability in individual responses. Identifying subgroups with differential treatment effects across multiple outcomes provides a more nuanced understanding of treatment efficacy. Existing methods for heterogeneous treatment effect (HTE) estimation often focus on a single endpoint, overlooking the potential of leveraging information from correlated secondary endpoints. In this work, we propose a multivariate probit-normal mixture model within the Seemingly Unrelated Regression Equations (SURE) framework to estimate and identify subgroup-specific treatment effects across multiple continuous endpoints. The model captures endpoint correlations and accounts for latent subgroups, where individuals may experience differential treatment effects. We evaluate the performance of the proposed approach using simulation studies and apply it to data from two clinical trials, demonstrating its potential to improve the precision of subgroup identification and enhance the interpretation of treatment effects across multiple endpoints.

A unified regression-based method for X-chromosome-inclusive Hardy–Weinberg equilibrium

♦Lin Zhang, Andrew Paterson, Lei Sun

Simon Fraser University, The Hospital for Sick Children, University of Toronto

How to best perform Hardy-Weinberg equilibrium (HWE) test for an X chromosomal SNP remains an open question, even using a sample of unrelated individuals. One simple strategy is to use female data only and apply the Pearson's chi-square test. Alternatively, earlier work has proposed a 2 df test that includes the deviation of observed male genotype counts from the expected based on pooled allele frequency estimate using both male and female data. Instead of the Pearson's chi-square test, we propose a new regression-based method that can (a) analyze both autosomal and X-chromosomal SNPs, (b) adjust for covariate effects if needed, (c) analyze related individuals, (d) include existing tests as special cases, and (e) lead to development of new tests. The proposed method builds from our recent robust allele-based (RA) regression method developed for conducting allelic association analysis. We illustrate the method by application to both phase 3 and high coverage whole genome sequence data from the 1000 genomes project.

25CHI051: Modern Statistical Learning

Statistical Learning via Partial Derivatives

♦Xiaowu Dai

University of California, Los Angeles

Traditional nonparametric estimation methods often lead to a slow convergence rate in large dimensions and require unrealistically large dataset sizes for reliable conclusions. We develop an approach based on partial derivatives, either observed or estimated, to effectively estimate the function at near-parametric convergence rates. This novel approach and computational algorithm could lead to methods useful to practitioners in many areas of science and engineering. Our theoretical results reveal behavior universal to this class of nonparametric estimation problems. We explore a general setting involving tensor product spaces and build upon the smoothing spline analysis of variance framework. For d-dimensional models under full interaction, the optimal rates with gradient information on p covariates are identical to those for the $(d-p)$-interaction models without gradients and, therefore, the models are immune to the curse of interaction. For additive models, the optimal rates using gradient information achieve the surprising parametric rate.

SAT: Data-light Uncertainty Set Merging via Synthetics, Aggregation, and Test Inversion

Shenghao Qin, Jianliang He, Bowen Gang, ♦Yin Xia

Fudan University, Yale University, Fudan University, Fudan University

The integration of uncertainty sets has diverse applications but also presents challenges, particularly when only initial sets and their control levels are available, along with potential dependencies. Examples include merging confidence sets from different distributed sites with communication constraints, as well as combining conformal pre- diction sets generated by different learning algorithms or data splits. In this article, we introduce an efficient and flexible Synthetic, Aggregation, and Test inversion (SAT) approach to merge various potentially dependent uncertainty sets into a single set. The proposed method constructs a novel class of synthetic test statistics, aggregates them, and then derives merged sets through test inversion. Our approach leverages the duality between set estimation and hypothesis testing, ensuring reliable coverage in dependent scenarios. The procedure is data-light, meaning it relies solely on initial sets and control levels without requiring raw data, and it adapts to any user-specified initial uncertainty sets, accommodating potentially varying coverage levels. Theoretical analyses and numerical experiments confirm that SAT provides finite-sample coverage guarantees and achieves small set sizes.

Optimal PhiBE — A New Framework for Continuous-Time Reinforcement Learning

♦Yuhua Zhu

University of California, Los Angeles

This talk addresses continuous-time reinforcement learning (RL) in settings where the system dynamics are governed by a stochastic differential equation but remains unknown, with only discrete-time observations available. While the optimal Bellman equation (optimal-BE) enables model-free algorithms, its discretization error is significant when the reward function oscillates. Conversely, model-based PDE approaches offer better accuracy but suffer from non-identifiable inverse problems.

To bridge this gap, we introduce Optimal-PhiBE, an equation that integrates discrete-time information into a PDE, combining the strengths of both RL and PDE formulations. Compared to the RL formulation, Optimal-PhiBE is less sensitive to reward oscillations, leading to smaller discretization errors. In linear-quadratic control, Optimal-PhiBE can even achieve accurate continuous-time optimal policy with only discrete-time information. Compared to the PDE formulation, it skips the identification of the dynamics and enables model-free algorithm derivation. Furthermore, we extend Optimal-PhiBE to higher orders, providing increasingly accurate approximations.

On the Optimality of Inference on the Mean Outcome under Optimal Treatment Regime

Shuoxun Xu, ♦Xinzhou Guo

UC Berkeley, HKUST

When an optimal treatment regime (OTR) is considered, we need to address the question of how good the OTR is in a valid and efficient way. The classical statistical inference applied to the mean outcome under the OTR, assuming the OTR is the same as the estimated OTR, might be biased when the regularity assumption that the OTR is unique is violated. Although several methods have been proposed to allow nonregularity in inference on the mean outcome under the OTR, the optimality of such inference is unclear due to challenges in deriving semiparametric efficiency bounds under potential nonregularity. In this paper, we address the bias issue induced by potential nonregularity via adaptive smoothing over the estimated OTR and develop a valid inference procedure on the mean outcome under the OTR regardless of whether the regularity assumption is satisfied or not. We establish the optimality of the proposed method by deriving a lower bound of the asymptotic variance for the robust asymptotically linear unbiased estimator to the mean outcome under the OTR and showing that our proposed estimator achieves the variance lower bound. The considered class of the estimator is general and includes the efficient regular estimator and the current state-of-the-art approach allowing nonregularity, and the derived lower bound of the asymptotic variance can be viewed as an extension of the classical semiparametric theory for OTR to a more general scenario allowing nonregularity. The merit of the proposed method is demonstrated by re-analyzing the ACTG 175 trial.

25CHI056: New advances in statistical theory, method and application

Analysis of Sparse Sufficient Dimension Reduction Models

Yeshan Withanage, Wei Lin, ♦Zhijian Li

JPMorgan Chase Bank, Ohio University, Beijing Normal - Hong Kong Baptist University

Since the introduction of the inverse regression method, sufficient dimension reduction (SDR) has become a very active topic in the literature. When the dimension p of x increases with the number of observations n, the traditional SDR methods may not perform well. In this study, we introduce how to apply the sparsity to the single-index models (a special SDR model) through Cumulative Mean Estimation (CUME) by using the LASSO approach and provide proof for the consistency of a variable selection procedure. We compare the performance of Lasso-CUME with Lasso-SIR (Sliced inverse regression) via extensive numerical studies.

Two-fold Varying-coefficient Mediation Models and Their Applications

Jie Xing, Le Zhou, Tiejun Tong, ♦WenWu Wang

Qufu Normal University, Hong Kong Baptist University, Hong Kong Baptist University, Qufu Normal University

Mediation analysis provides a technical approach to reveal the complex relationships between variables, and researchers can use causal mediation analysis to distinguish more targeted interventions for different individuals and different scenario conditions. In real life, the strength of the effects between the variables in a causal relationship is often influenced by many other variables. Meanwhile, there are still relatively few studies on the theories and applications of the multiple varying-coefficient mediation models. In this paper, we study two influence mechanisms of two mediators in the two-fold varying-coefficient mediation model. In the first case, the mediators are in a parallel relationship; and in the second case, the mediators are sequential, where one mediator has a fixed coefficient and the other has a varying coefficient. Following these two cases, we propose the parallel two-fold varying-coefficient mediation model and the chain two-fold varying-coefficient mediation model, and regard the coefficient functions as smooth functions of the effect modifier. We further derive the spline estimators of the direct and indirect effects, as well as establish the asymptotic normality of all or part of the estimators. Simulation studies show that our new models and methods for the direct and indirect effects perform well, and they are also able to capture the dynamic changes of the coefficient functions, in the two real data examples, compared to the classic mediation model with constant coefficients.

When Tukey meets Chauvenet: a new boxplot criterion for outlier detection

♦Hongmei Lin, Riquan Zhang, Tiejun Tong

Shanghai University of International Business and Economics, Shanghai University of International Business and Economics, Hong Kong Baptist University

The box-and-whisker plot, introduced by Tukey (1977), is one of the most popular graphical methods in descriptive statistics.

On the other hand, however, Tukey's boxplot is free of sample size, yielding the so-called ``one-size-fits-all'' fences for outlier detection.

Although improvements on the sample size adjusted boxplots do exist in the literature, most of them are either

not easy to implement or lack justification. As another common rule for outlier detection, Chauvenet's criterion uses the sample mean and standard derivation to perform the test, but it is often sensitive to the included outliers and hence is not robust. In this paper, by combining Tukey's boxplot and Chauvenet's criterion, we introduce a new boxplot, namely the Chauvenet-type boxplot, with the fence coefficient determined by an exact control of the outside rate per observation. Our new outlier criterion not only maintains the simplicity of the boxplot from a practical perspective, but also serves as a robust Chauvenet's criterion. Simulation study and a real data analysis on the civil service pay adjustment in Hong Kong demonstrate that the Chauvenet-type boxplot performs extremely well regardless of the sample size, and can therefore be highly recommended for practical use to replace both Tukey's boxplot and Chauvenet's criterion. Lastly, we also develop a new R package `C.boxplot' that can be implemented very user-friendly, and moreover make the source files and example codes freely available.

Structural Testing of High-dimensional Correlation Matrices

♦Tingting Zou, Guangren Yang, Ruitao Lin, Guo-Liang Tian, Shurong Zheng,

Jilin University, Jinan University, The University of Texas MD Anderson Cancer Center, Southern University of Science and Technology, Northeast Normal University,

Due to scale invariance, correlation matrices play a critical role in multivariate statistical analysis. Statistical inference about correlation matrices encounter enormous challenges and is fundamentally different from inference about covariance matrices in both low- and high-dimensional settings. This paper studies the test of general linear structures of high-dimensional correlation matrices, which include commonly-used banded matrices and compound symmetry matrices as special cases. We first propose a procedure using the quadratic loss function to estimate the unknown parameters associated with the assumed linear structure. We then develop test statistics, based on the quadratic and infinite norms, which are suitable for dense and sparse alternatives, respectively. The limiting distributions of our proposed test statistics are derived under the null and alternative hypotheses. Extensive simulation studies are conducted to demonstrate the finite sample performance of our proposed tests. Moreover, a real data example is provided to show the applicability and the practical utility of the tests.

25CHI057: New frontiers in large-scale data analysis with applications to heterogeneous data

Kernel Regression Utilizing External Information as Constraints

♦Chi-Shian Dai, Shao Jun

National Cheng Kung Univeristy, University of Wisconsin-Madison

In modern scientific research and practice, the widespread availability of large and varied data sources has shifted the analytical focus from single data sets to integrating multiple data sets. Effectively merging these diverse resources requires addressing critical data-quality challenges, including incomplete or partial covariates, summary-level information, and heterogeneity across different data sets. Overcoming these obstacles is vital for achieving reliable and comprehensive data integration.

In this talk, I will introduce a novel methodology that leverages external information as a constraint in kernel regression models. This framework significantly enhances predictive performance by incorporating partial covariate information, accommodating summary-level data, and managing heterogeneous data sets. This approach paves the way for more robust, accurate, and versatile data integration in various scientific and applied settings.

An alternative measure for quantifying the heterogeneity in meta-analysis

Ke Yang, Enxuan Lin, Wangli Xu, Liping Zhu, ♦Tiejun Tong

Beijing University of Technology, Innovent Biologics, Inc., Renmin University of China, Renmin University of China, Hong Kong Baptist University,

Quantifying the heterogeneity is an important issue in meta-analysis, and among the existing measures, the I^2 statistic is most commonly used. In this paper, we first illustrate with a simple example that the I^2 statistic is heavily dependent on the study sample sizes, mainly because it is used to quantify the heterogeneity between the observed effect sizes. To reduce the influence of sample sizes, we introduce an alternative measure that aims to directly measure the heterogeneity between the study populations involved in the meta-analysis. We further propose a new estimator, namely the I_A^2 statistic, to estimate the newly defined measure of heterogeneity. For practical implementation, the exact formulas of the I_A^2 statistic are also derived under two common scenarios with the effect size as the mean difference (MD) or the standardized mean difference (SMD). Simulations and real data analyses demonstrate that the I_A^2 statistic provides an asymptotically unbiased estimator for the absolute heterogeneity between the study populations, and it is also independent of the study sample sizes as expected. To conclude, our newly defined I_A^2 statistic can be used as a supplemental measure of heterogeneity to monitor the situations where the study effect sizes are indeed similar with little biological difference. In such scenario, the fixed-effect model can be appropriate; nevertheless, when the sample sizes are sufficiently large, the I^2 statistic may still increase to 1 and subsequently suggest the random-effects model for meta-analysis.

Statistical inference for large-scale multi-source heterogeneous data

Jiuzhou Miao, ♦Li Cai, Suojin Wang

Zhejiang Gongshang university, Zhejiang Gongshang University, Texas A&M university

In the era of digital information, people are faced with data that may not only be large-scale but also heterogeneous. In this paper, we study statistical inference for the overall population mean function of large-scale multi-source heterogeneous datasets. By borrowing hierarchical sampling methods and divide-and-conquer techniques, we propose a weighted local linear estimator for the overall population mean function of multi-source heterogeneous data. Through studying the pointwise convergence properties and extreme value distribution properties of the estimator, we construct asymptotically accurate simultaneous confidence bands and pointwise confidence intervals for large-scale multi-source heterogeneous data. Our proposed methods are applicable not only to scenarios of heterogeneous data but also to scenarios of homogeneous data using divide-and-conquer methods. Numerical simulation studies show that the proposed methods perform well in analyzing both large-scale multi-source heterogeneous data and homogeneous data. As an illustration, we apply the proposed methods to hypothesis testing problems on Beijing multi-site air-quality data and U.S. census data.

Power Enhancement Subsampling Adaptive Ensemble Test

♦Xuehu Zhu

Xi'an Jiaotong University

Specification testing constitutes a critical element within the framework of statistical methodologies. However, many prevalent tests suffer from high computational complexity, especially when dealing with large datasets, and are inefficient in resource-constrained environments. To address this challenge, we propose a Subsampling Adaptive-to-Model Ensemble Test (SAET), which adeptly integrates information-based subsampling techniques with adaptive-to-model hybrid tests based on moment and conditional moment methodologies. SAET integrates the advantages of both local and global smoothing tests, while mitigating their drawbacks, and can enhance the power of tests compared to traditional uniform subsampling methods, all while maintaining computational efficiency. To further optimize the utility of selected samples, we develop the Enhanced Subsampling Adaptive-to-Model Ensemble Test (ESAET), which enhances the power of SAET. The two methods are readily applicable to various types of large sample datasets with limited budget and can be extended to the construction of other ensemble-based testing frameworks. Extensive numerical studies conducted on both simulated and real-world datasets demonstrate the superior performance and robustness of the two tests.

25CHI062: On Statistical Stability and Ensemble Learning

Integrating Inference Results via Synthetic Statistics.

Shenghao Qin, Jianliang He, ♦Bowen Gang, Yin Xia

Fudan University, Yale University, Fudan University, Fudan University

Learning from the collective wisdom of crowds parallels the statistical concept of fusion learning from multiple data sources or studies. However, integrating inferences from diverse sources poses significant challenges due to cross-source heterogeneity and data-sharing limitations. Studies often rely on varied designs and modeling techniques, and stringent data privacy norms can prohibit even the sharing of summary statistics. In this talk, I will discuss the construction of "synthetic statistics" that mimic the summary statistics used for inference, enabling the fusion of inference results from multiple sources.

A Gaussian Stability framework for Post-Selection Inference

Jiajun Sun, Chendi Wang, ♦Wei Zhong

Xiamen University, Xiamen University, Xiamen University

When we use data for statistical inference after model selection with the same set of data, the classical theory of statistical inference may not be valid. Zrnic and Jordan (2023) proposed a method to resolve this issue through algorithmic stability, introducing randomization to enable post-selection corrections without computational burdens. However, the previous approach has the drawback of being overly conservative when applied to complex composite algorithms because of the shortcomings of the stability framework and the heavy tail nature of the Laplace noise they use. To this end, we propose a Gaussian stability framework for Post-Selection Inference (GPS), which has better algorithmic stability as well as narrower confidence intervals. Theoretically, we establish the coverage guarantee of the proposed method and examine the regimes under which the proposed method yields a narrower confidence interval width compared to previous approaches. In addition, to reduce randomness in inferring irrelevant parts without losing the validity of statistical inference, we propose an aggregation method. Numerical studies demonstrate that the proposed method has superior empirical performance, thereby offering a more robust and practical solution for post-selection inference.

A simple and powerful method for large-scale composite null hypothesis testing with applications in mediation analysis

♦*Yaowu Liu*

Southwestern University of Finance and Economics

Large-scale mediation analysis has received increasing interest in recent years, especially in genome-wide epigenetic studies. The statistical problem in large-scale mediation analysis concerns testing composite null hypotheses in the context of large-scale multiple testing. The classical Sobel's and joint significance tests are overly conservative and therefore are underpowered in practice. In this work, we propose a testing method for large-scale composite null hypothesis testing to properly control the type I error and hence improve the testing power. Our method is simple and essentially only requires counting the number of observed test statistics in a certain region. Non-asymptotic theories are established under weak assumptions and indicate that the proposed method controls the type I error well and

is powerful. Extensive simulation studies confirm our non-asymptotic theories and show that the proposed method controls the type I error in all settings and has strong power. A data analysis on DNA methylation is also presented to illustrate our method.

25CHI066: Recent Advances in Correcting Measurement Error in Epidemiological Research

Causal inference for the effects of mismeasured covariates through double/debiased machine learning

Gang Xu, Xin Zhou, Molin Wang, Boya Zhang, Donna Spiegelman, ♦*Zuoheng Wang*

Yale University, Yale University, Harvard University, Harvard University, Yale University, Yale University

One way to quantify exposure to air pollution and its constituents in epidemiologic studies is to use an individual's nearest monitor. This strategy results in potential inaccuracy in the actual personal exposure, introducing bias in estimating the health effects of air pollution and its constituents, especially when evaluating the causal effects of correlated multi-pollutant constituents measured with correlated error. This paper addresses estimation and inference for the causal effect of one constituent in the presence of other PM2.5 constituents, accounting for measurement error and correlations. We used a linear regression calibration model in an external validation study, and extended a double/debiased machine learning (DML) approach to correct for measurement error and estimate the effect of interest in the main study. We demonstrated that the DML estimator with regression calibration is consistent and derived its asymptotic variance. Simulations showed that the proposed estimator reduced bias and attained nominal coverage probability across most simulation settings. We applied this method to assess the causal effects of PM2.5 constituents on cognitive function in the Nurses' Health Study and identified two PM2.5 constituents, Br and Mn, that showed a negative causal effect on cognitive function after measurement error correction.

Time-to-Event Analysis of Preterm Birth Accounting for Gestational Age Uncertainties

♦*Yuzi Zhang, Joshua Warren, Hua Hao, Howard Chang*

Division of Biostatistics, The Ohio State University, Department of Biostatistics, Yale University, Department of Environmental Health, The Rollins School of Public Health of Emory University, Department of Biostatistics and Bioinformatics, Emory University

A time-to-event analysis is advocated for examining associations between time-varying environmental exposures and preterm birth in cohort studies. While the identification of preterm birth entirely depends on gestational age, the true gestational age is rarely known in practice. Obstetric estimate (OE) and gestational age based on the date of last menstrual period (LMP) are two commonly used measurements, but both suffer from various sources of error. Uncertainties in gestational age result in both outcome misclassification and measurement error of time-varying exposures which can potentially introduce serious bias in health effect estimates. Motivated by the lack of validation data in large population-based studies, we develop a hierarchical Bayesian model that utilizes the two error-prone gestational age estimates to examine time-varying exposures on the risk of preterm birth while accounting for uncertainties in the estimates. The proposed approach introduces two discrete-time hazard models for the latent true gestational ages that are preterm (<37 weeks) or term (≥ 37 weeks). Then two multinomial models are adopted for characterizing misclassifications resulting from using OE-based and LMP-based gestational age. The proposed modeling framework permits the joint estimation of preterm birth risk factors and parameters characterizing gestational age misclassifications without validation data. We apply the proposed method to a birth cohort based on birth records from Kansas in 2010. Our analysis finds robust positive associations between exposure to ozone during the third trimester of pregnancy and preterm birth even after accounting for gestational age uncertainty.

Survival analysis adjusting for measurement error in a cumulative exposure variable: radon progeny to lung cancer mortality

♦*Molin Wang, Yue Yang, Donna Spiegelman*

Harvard University, Harvard University, Yale University

Exposure measurement error is a common occurrence in various epidemiologic fields, with radiation epidemiology at the top of the list. Failure to properly assess and adjust for uncertainties in radiation dosimetry could lead to biased effect estimates. Moreover, characterizations of health impacts obtained without countering error in exposure levels could potentially misinform policy makers, when they are, for example, setting the radiation safety levels in occupational and residential settings referencing unadjusted dose-response relationships between error-prone radiation levels and observed adverse health outcomes. Therefore, from both the statistical advancement and public health policy perspectives, it is of great importance to develop and discuss statistical methods in countering the influences of such exposure measurement error and providing valid health outcome effects into the policy decision pipeline. In this talk, I will present statistical methods for estimating exposure-outcome associations, adjusting for the exposure measurement errors when the exposure takes the form of a cumulative total, in the settings of Cox proportional hazard models and excess relative risk models under both the additive and multiplicative measurement errors. The proposed methods will be illustrated using data from the field of radiational epidemiology.

Generalized Methods-of-Moments estimation and inference for the assessment of multiple imperfect measures of physical activity in validation studies

Zexiang Li, Donna Spiegelman, Molin Wang, Zuoheng Wang, ♦Xin Zhou

Yale School of Public Health, Yale School of Public Health, Harvard T.H. Chan School of Public Health, Yale School of Public Health, Yale School of Public Health,

Assessing diet and physical activity in free-living populations is prone to substantial measurement error. To evaluate and improve the quality of these measurements, their correlations with corresponding true exposures play a critical role in informing the iterative refinement of measures. In practice, a major barrier is that the gold standard for the true exposure does not exist for many cases, instead, a reference measurement that contains some errors is available. One widely used method for this issue is the Method of Triads (MOT), which involves biomarkers to obtain the correlation of each measurement with the unobserved truth by conducting a triangular comparison. However, MOT no longer produces consistent estimates when the errors in the measurements are correlated with each other. In this work, under the assumption that the moments are correct, we develop semi-parametric generalized method of moments estimators for correlations and other quantities of interest, including the de-attenuation factor and intra-class correlation coefficients characterizing the random within-person variation around each measurement. Unlike standard MOT, our method allows correlated errors in assessment measures and incorporates Wald statistics to determine which errors across measurement types are uncorrelated. A robust variance is derived to enable asymptotic inference. The performance of the proposed method has been evaluated in extensive simulation studies and data analysis in the Men's Lifestyle Validation Study.

25CHI067: Recent advances in high dimensional data and machine learning

Privacy-Preserving Transfer Learning for Community Detection using Locally Distributed Multiple Networks

Xiao Guo, Xuming He, Xiangyu Chang, ♦Shujie Ma

Northwest University of China, Washington University in St. Louis, Xi'an Jiaotong University, University of California-Riverside

In this talk, I will introduce a new spectral clustering-based method called TransNet for transfer learning in community detection of network data. Our goal is to improve the clustering performance of the target network using auxiliary source networks, which are locally stored across various sources, privacy-preserved, and heterogeneous. Notably, we allow the source networks to have distinct privacy-preserving and heterogeneity levels that often happen in practice. To better utilize the information from the heterogeneous and privacy-preserved source networks, we propose a new adaptive weighting method to aggregate the eigenspaces of the source networks and a regularization method that can automatically combine the weighted average eigenspace of the source networks with the eigenspace of the target network to achieve an optimal balance between them. We also demonstrate that TransNet performs better than both the estimator only using the target network and the estimator using the weighted source networks.

Statistical inference for high-dimensional convoluted rank regression

Leheng Cai, ♦Xu Guo, Heng Lian, Liping Zhu

Tsinghua University, Beijing Normal University, City University of Hong Kong, Renmin University of China

High-dimensional penalized rank regression is a powerful tool for modeling high-dimensional data due to its robustness and estimation efficiency. However, the non-smoothness of the rank loss brings great challenges to the computation. To solve this critical issue, high-dimensional convoluted rank regression has been recently proposed, introducing penalized convoluted rank regression estimators. However, these developed estimators cannot be directly used to make inference. In this paper, we investigate the statistical inference problem of high-dimensional convoluted rank regression. The use of U-statistic in convoluted rank loss function presents challenges for the analysis. We begin by establishing estimation error bounds of the penalized convoluted rank regression estimators under weaker conditions on the predictors. Building on this, we further introduce a debiased estimator and provide its Bahadur representation. Subsequently, a high-dimensional Gaussian approximation for the maximum deviation of the debiased estimator is derived, which allows us to construct simultaneous confidence intervals. For implementation, a novel bootstrap procedure is proposed and its theoretical validity is also established. Finally, simulation and real data analysis are conducted to illustrate the merits of our proposed methods.

Clustering functional data with measurement errors: a simulation-based approach

Tingyu Zhu, ♦Lan Xue, Carmen Tekwe, Keith Diaz, Mark Benden, Roger Zoh

Oregon State University, Oregon State University, Indiana University, Columbia University Medical Center, Texas A&M University, Indiana University

Clustering analysis of functional data, which comprises observations that evolve continuously over time or space, has gained increasing attention across various scientific disciplines. Practical applications often involve functional data that are contaminated with measurement errors arising from imprecise instruments, sampling errors, or other sources. These errors can significantly distort the inherent data structure, resulting in erroneous clustering outcomes. In this paper, we propose a simulation-based approach designed to mitigate the impact of measurement errors. Our proposed method estimates the distribution of functional measurement errors through repeated measurements. Subsequently, the clustering algorithm is applied to simulated data generated from the conditional distribution of the unobserved true functional data given the observed contaminated functional data, accounting for the adjustments made to rectify measurement errors. We illustrate through simulations show that the proposed method has improved numerical performance than the naive methods that neglect such errors. Our proposed method was applied to a childhood obesity study, giving more reliable clustering results.

Fair classification with continuous sensitive attribute

♦*Xianli Zeng, Edgar Dobriban*

Xiamen University

This paper addresses the problem of fair classification with a continuous protected attribute, a crucial yet underexplored topic.

Most work in fair machine learning focuses on discrete protected attributes, and may lead to biases when used for continuous attributes such as income. To develop methods suitable for continuous attributes, we first define suitable notions of disparity by extend the concept of ``linear disparity" from discrete features to continuous ones. We then derive the form of the maximally accurate (Bayes-optimal) classifier controlling such a disparity metric at the level of the full population from which the data is sampled. To estimate this fair Bayes-optimal classifier based on the observed data, we propose a novel post-processing method called FairBayes-NBS. We provide theoretical results on the asymptotic convergence rate of FairBayes-NBS. Simulations and experiments on the standard AdultCensus dataset suggest that our methods perform well in a broad range of settings.

25CHI068: Recent Advances in Machine Learning Techniques for Point Process Models

Score Matching for Point Processes

♦*Feng Zhou*

Renmin University of China

Score matching estimators have gained widespread attention in recent years partly because they are free from calculating the integral of normalizing constant, thereby addressing the computational challenges in maximum likelihood estimation (MLE). Some existing works have proposed score matching estimators for point processes. However, this work demonstrates that the incompleteness of the estimators proposed in those works renders them applicable only to specific problems, and they fail for more general point processes. To address this issue, this work introduces the weighted score matching estimator to point processes. Theoretically, we prove the consistency of our estimator and establish its rate of convergence. Experimental results indicate that our estimator accurately estimates model parameters on synthetic data and yields results consistent with MLE on real data. In contrast, existing score matching estimators fail to perform effectively.

Bayesian inference for independent cluster point processes

Marie-Colette van Lieshout, ♦*Changqing Lu*

National Research Institute for Mathematics and Computer Science in the Netherlands; University of Twente, National Research Institute for Mathematics and Computer Science in the Netherlands; University of Twente

Spatial and spatio-temporal cluster point processes have been extensively studied, with much of the literature assuming Poisson structures for parent, child, or both processes due to their straightforward probability density functions. Frequentist and Bayesian approaches have been developed to infer such models. However, in practice, assuming a Poisson parent process often results in an overabundance of clusters, while assuming a Poisson child process may not adequately capture real-world complexities, as illustrated by the arson fire data example. In this paper, we propose a two-step Bayesian inference framework for independent cluster point processes that are not restricted by Poisson assumptions. Specifically for the fire data, we introduce a model incorporating a repulsive prior and a shifted-Poisson Gaussian child. We develop a Markov chain Monte Carlo method to estimate cluster states given fixed model parameters and a Monte Carlo Expectation-maximization method to estimate the model parameters themselves. We show the convergence of the proposed approach and validate its effectiveness through a simulation study and the application to the real fire data. The Bayesian framework we present is flexible and can accommodate a variety of parent and child processes provided that their probability functions have an analytical form. Furthermore, from a practical perspective, our approach enables to predict the probability of being-forming clusters in a spatio-temporal manner.

Residual TPP: A unified lightweight approach for event stream data analysis

Ruoxin Yuan, ♦*Guanhua Fang*

Fudan University, Fudan University

This work introduces Residual TPP, a novel, unified, and lightweight approach for analyzing event stream data. It leverages the strengths of both simple statistical TPPs and expressive neural TPPs to achieve superior performance. Specifically, we propose the Residual Events Decomposition (RED) technique in temporal point processes, which defines a weight function to quantify how well the intensity function captures the event characteristics. The RED serves as a flexible, plug-and-play module that can be integrated with any TPP model in a wide range of tasks. It enables the identification of events for which the intensity function provides a poor fit, referred to as residual events. By combining RED with a Hawkes process, we capture the self-exciting nature of the data and identify residual events. Then an arbitrary neural TPP is employed to take care of residual events. Extensive experimental results demonstrate that Residual TPP consistently achieves state-of-the-art goodness-of-fit and prediction performance in multiple domains and offers significant computational advantages as well.

Tensor-based Estimation and Inference for High-Dimensional Multivariate Point Process

Xiwei Tang, ♦*Gannggang Xu, Jingfei Zhang, Yongtao Guan*

University of Texas at Dallas, University of Miami, Emory University, Chinese University of Hong Kong, Shenzhen

We investigate the complex dependency structures among a diverging number of point processes by leveraging a low-rank tensor decomposition framework. This approach enables scalable modeling of high-dimensional event data while capturing both temporal and cross-process interactions in a parsimonious manner. We develop estimators based on this framework and establish their theoretical properties, including consistency and convergence rates, under suitable regularity conditions. To assess the practical utility of the proposed method, we conduct extensive simulation studies under a variety of settings and demonstrate its advantages in terms of estimation accuracy and inference validity. We further apply our method to a real-world dataset to illustrate its effectiveness in uncovering meaningful patterns in high-dimensional point process data.

25CHI074: Recent Advances in Statistical and Machine Learning

Contextual Dynamic Pricing: Algorithms, Optimality, and Local Differential Privacy Constraints

Zifeng Zhao, ♦Feiyu Jiang, Yi Yu

University of Notre Dame, Fudan University, University of Warwick

We study contextual dynamic pricing problems where a firm sells products to T sequentially-arriving consumers, behaving according to an unknown demand model. The firm aims to minimize its regret over a clairvoyant that knows the model in advance. The demand follows a generalized linear model (GLM), allowing for stochastic feature vectors in R^d encoding product and consumer information. We first show the optimal regret is of order \sqrt{dT}, up to logarithmic factors, improving existing upper bounds by a \sqrt{d} factor. This optimal rate is materialized by two algorithms: a confidence bound-type algorithm and an explore-then-commit (ETC) algorithm. A key insight is an intrinsic connection between dynamic pricing and contextual multi-armed bandit problems with many arms with a careful discretization.

We further study contextual dynamic pricing under local differential privacy~(LDP) constraints. We propose a stochastic gradient descent-based ETC algorithm achieving regret upper bounds of order d\sqrt{T}/\epsilon, up to logarithmic factors, where \epsilon>0 is the privacy parameter. The upper bounds with and without LDP constraints are matched by newly constructed minimax lower bounds, characterizing costs of privacy. Moreover, we extend our study to dynamic pricing under mixed privacy constraints, improving the privacy-utility tradeoff by leveraging public data. This is the first time such setting is studied in the dynamic pricing literature and our theoretical results seamlessly bridge dynamic pricing with and without LDP. Extensive numerical experiments and real data applications are conducted to illustrate the efficiency and practical value of our algorithms.

Sequential knockoffs for variable selection in reinforcement learning

Tao Ma, ♦Jin Zhu, Hengrui Cai, Zhengling Qi, Yunxiao Chen, Chengchun Shi, Eric Laber

London School of Economics and Political Science, London School of Economics and Political Science, University of California, Irvine, George Washington University, London School of Economics and Political Science, London School of Economics and Political Science, Duke University

In real-world applications of reinforcement learning, it is often challenging to obtain a state representation that is parsimonious and satisfies the Markov property without prior knowledge. Consequently, it is common practice to construct a state larger than necessary, e.g., by concatenating measurements over contiguous time points. However, needlessly increasing the dimension of the state may slow learning and obfuscate the learned policy. We introduce the notion of a minimal sufficient state in a Markov decision process (MDP) as the subvector of the original state under which the process remains an MDP and shares the same reward function as the original process. We propose a novel SEquEntial Knockoffs (SEEK) algorithm that estimates the minimal sufficient state in a system with high-dimensional complex nonlinear dynamics. In large samples, the proposed method achieves selection consistency. As the method is agnostic to the reinforcement learning algorithm being applied, it benefits downstream tasks such as policy learning. Empirical experiments verify theoretical results and show the proposed approach outperforms several competing methods regarding variable selection accuracy and regret.

Joint robust estimation

♦Lihu Xu

University of Macau

In this talk, we shall study a joint robust estimation for the heavy tailed data which only have finite (3+\epsilon)-th moment , which can estimate the target parameters and the variance simultaneously. We shall apply this estimation to three cases: robust mean estimation, robust linear regression, high dimensional robust linear regression.

Consistent Selection of the Number of Groups in Panel Models via Cross-Validation

Zhe Li, Changliang Zou, ♦Xuening Zhu

Fudan University

Group number selection is a key problem for group panel data modeling. In this work, we develop a cross-validation (CV) method to tackle this problem. Specifically, we split the panel data into two data folds on the time span, with group structure preserved for individuals. We first estimate the group memberships and parameters on one data fold, then we plug in the estimates and utilize the other data fold to evaluate a designed criterion. Subsequently, the group number is estimated by minimizing the average criterion across all data folds. The proposed CV method has two advantages compared to existing approaches. First, the method is totally data-driven, thus no further tuning parameters are involved. Second, the method can be flexibly applied to a wide range of panel data models. Theoretically, we establish the estimation consistency by taking advantage of the optimization property of the estimation algorithm. Experiments are carried out with a variety of synthetic datasets and panel models to further illustrate the advantages of the proposed method. Lastly, the CV method is employed to analyze the heterogeneous patterns of stock volatilities in the Chinese stock market through the financial crisis.

25CHI085: Recent developments of high dimensional model checking

Testing mutually exclusive hypotheses for multi-response regressions

♦Jiaqi Huang, Wenbiao Zhao, Lixing Zhu

Beijing Normal University, China University of Mining & Technology, Beijing, Beijing Normal University

This paper proposes an adaptive-to-model test to check the null hypothesis with no more than one coordinate of the response vector relating to the predictor vector in parametric multi-response regressions. To this end, we decompose the null hypothesis into several mutually exclusive sub-null hypotheses and suggest a model identification to construct an adaptive-to-sub-null hypothesis test tackling their mutual exclusiveness, and an adaptive-to-regression test handling the regression function mis-specification. The final test combines a further model identification to be an adaptive-to-model hybrid of these two tests. It has the chi-square weak limit under the null hypothesis even when the dimensions of the response and the predictor vectors increase with the sample size and is omnibus. We conduct a systematic analysis of the significance level maintenance and power performance of the test to reveal its different sensitivity rates of convergence to different sub-local alternatives distinct from the null hypothesis. This is a significant distinction against any existing model checking problems for regressions. Further, the proposed model identifications can also assist in identifying the responses with non-constant regressions and testing their mis-specification. Numerical studies include simulations to examine the finite sample performances and to illustrate real data analyses for two data sets.

Goodness-of-Fit Tests for High-Dimensional Regression Models via Projections

Wen Chen, Jie Liu, Heng Peng, ♦Falong Tan, Lixing Zhu

Hunan University, Hunan University, Hong Kong Baptist University, Hunan University, Beijing normal University

In this talk, we proposed a new method for testing the goodness of fit for high dimensional generalized linear regression models when the number of covariates may be much larger than the sample size. Most existing model checking methods in the literature does not work for high dimension regression models as they suffer from the curse of dimensionality and rely on the asymptotic linearity and normality of the estimator of the parameters. Our method is based on random projections which largely avoid the “curse of dimensionality”. Further, our test only need the convergence rate of the estimators of the high dimensional parameters and does not rely on the asymptotic expansion or the normality of these estimators. The asymptotic properties of the test statistics are investigated under the null and the local and global alternatives when the number of covariates is much larger than the sample sizes. We further proposed a combination method to enhance the power performance of the tests. Detailed simulation studies and a real data analysis are conducted to illustrate the effectiveness of our methodology.

Model checking for parametric regressions in transfer learning

Chuhan Wang, Jiaqi Huang, ♦Xuerui Li

Beijing Normal University, Beijing Normal University, Beijing Normal University

This paper investigates whether the parametric regression model is specified correctly in both the source and target data and whether the regression pattern for the source is consistent with that of the target. This is crucial before employing methods designed for model-based transfer learning under covariate shift assumptions. Neither classical model checking for regressions nor two-sample tests for regressions can achieve this. To this end, we propose a novel adaptive-to-regression test statistic that is asymptotically distribution-free. This test has a chi-square weak limit under the null hypothesis, thereby maintaining the significance level and enabling the determination of the critical value without the need for resampling techniques. We also conduct a systematic analysis of the test's power performance to reveal its varying sensitivity rates of convergence to different sub-local alternatives distinct from the null hypothesis.

A Chi-Square Specification Test with One-Class Support Vectors

Yuhao Li, ♦Xiaojun Song

Xi’an Jiaotong-Liverpool University, Peking University

Specification tests are commonly employed in practice to assess whether a model is valid or invalid. Important examples include the Integrated Conditional Moment and Kernel Conditional Moment tests. Despite their theoretical advantages, they are nonpivotal and computationally intensive, requiring quadratic time and bootstrap procedures for critical value determination. This paper proposes a novel approach to address these issues by projecting the mean difference element into a direction of the Reproducing Kernel Hilbert Space (RKHS) using a real-valued analytic kernel. The projection direction is determined by a finite set of location points and corresponding weights. The test statistic is linear in time, omnibus, and converges to the standard chi-square distribution under the null hypothesis. All asymptotic results are obtained using basic limiting theorems. The One-Class Support Vector Machine (OCSVM) algorithm is employed to select the finite location points and weights, enhancing the test’s power. Simulation studies demonstrate that the proposed test exhibits favorable finite sample properties.

25CHI086: Recent Progresses in Nonparametric and Semiparametric Statistics

Enhanced Fused Sufficient Representation Learning for Neuroimaging Data

Yue Chen, Baiguo An, Linglong Kong, Xueqin Wang, ♦Wenliang Pan

Capital University of Economics and Business, Capital University of Economics and Business, University of Alberta, University of Science and Technology of China, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Neuroimaging data analysis presents significant challenges due to the high-dimensional, complex, and spatially structured nature of the data. Effective representation learning for neuroimaging must not only capture predictive relationships but also preserve spatial and anatomical context to ensure interpretability for clinical applications. However, most existing methods overlook these critical aspects, resulting in representations that fail to fully utilize structural information and lack clinical relevance. To address these limitations, we propose a novel approach, Enhanced Fused Sufficient Representation Learning (EFSRL), which integrates sufficient representation learning with region detection. Our method's core is HSIC-FUSE, a measure aggregating normalized Hilbert-Schmidt Independence Criterion (HSIC) values across multiple kernels to promote sufficient representation without relying on arbitrary kernel selection. These ensure both robustness and interpretability, which are essential for clinical tasks. We also introduce a dual-network architecture that alternates between learning representations and selecting key regions, facilitating more accurate and meaningful interpretations. Through extensive experiments on synthetic and real-world medical imaging data, including the ADNI dataset, we demonstrate that EFSRL outperforms existing methods. Our approach generates interpretable representations tailored for various medical imaging tasks, highlighting its potential for practical applications in healthcare.

Unsupervised optimal deep transfer learning for classification under general conditional shift

Junjun Lang, ♦Yukun Liu

East China Normal University, East China Normal University

Classifiers trained solely on labeled source data may yield misleading results when applied to unlabeled target data drawn from a different distribution. Transfer learning can rectify this by transferring knowledge from source to target data, but its effectiveness frequently relies on stringent assumptions, such as label shift. In this paper, we introduce a novel General Conditional Shift (GCS) assumption, which encompasses label shift as a special scenario. Under GCS, we demonstrate that both the target distribution and the shift function are identifiable under mild conditions. To estimate the conditional probabilities ${\bm\eta}_P$ for source data, we propose leveraging deep neural networks (DNNs). Subsequent to transferring the DNN estimator, we estimate the target label distribution ${\bm\pi}_Q$ utilizing a pseudo-maximum likelihood approach. Ultimately, by incorporating these estimates and circumventing the need to estimate the shift function, we construct our proposed Bayes classifier. We establish concentration bounds for our estimators of both {\bm\eta}_P$ and ${\bm\pi}_Q$ in terms of the intrinsic dimension of ${\bm\eta}_P$. Notably, our DNN-based classifier achieves the optimal minimax rate, up to a logarithmic factor. A key advantage of our method is its capacity to effectively combat the curse of dimensionality when ${\bm\eta}_P$ exhibits a low-dimensional structure. Numerical simulations, along with an analysis of an Alzheimer's disease dataset, underscore its exceptional performance.

Semi-parametric inference on inequality measures with non-ignorable non-response using callback data

♦Chunlin Wang

Xiamen University

Measuring income inequality is vital in economics and official statistics. In practice, the household survey data sets that are typically used to measure income inequality inevitably suffer from non-ignorable non-response, that is, the response probabilities depend on the missing actual income values. Existing methods assuming fully observed income data may not be applicable or lead to distorted statistical inference based on the complete-case analysis. In this paper, we consider efficient and reliable estimation and inference of popular measures of inequality in the presence of non-ignorable non-response. We exploit the callback data routinely collected along with many surveys to tackle the model identifiability issue and correct the biased sample through a semi-parametric modeling strategy that does not require any parametric specification on the income distribution. A full likelihood approach is developed to estimate various inequality measures. To circumvent complex numerical optimization, we further devise a novel expectation-maximization algorithm for stable and convenient computation. The asymptotic properties of the proposed estimators are established, which enable valid statistical inference of the inequality measures. The simulation results demonstrate that the proposed semi-parametric method corrects the non-response bias of the estimated inequality measures, is robust to income distributions, and leads to efficient inference results. We apply the proposed inference procedures of inequality measures to a real income survey data set with non-ignorable non-response for illustration.

Matching-based Policy Learning

♦Ying Yan

Sun Yat-sen University

In this talk, we propose a matching-based policy learning framework. We adapt standard and bias-corrected matching methods to estimate an alternative form of the value function: the advantage function, which can be interpreted as the expected improvement achieved by implementing a given policy compared to the equiprobable random policy. We then learn the optimal policy over a restricted policy class by maximizing the matching estimator of the advantage function. We derive a non-asymptotic high probability bound for the regret of the learned optimal policy, and show that the learned policy is almost rate-optimal. The competitive finite sample performance of the proposed method compared to weighting-based and outcome modeling-based learning methods is demonstrated in extensive simulation studies and a real data application.

25CHI090: Scalable Learning and Knowledge Transfer for Complex Biomedical Data

Data-Driven Knowledge Transfer in Batch Q* Learning

Elynn Chen, Xi Chen, ♦Wenbo Jing

New York University, New York University, New York University

In data-driven decision-making across marketing, healthcare, and education, leveraging large datasets from existing ventures is crucial for navigating high-dimensional feature spaces and addressing data scarcity in new ventures. We investigate knowledge transfer in dynamic decision-making by focusing on batch stationary environments and formally defining task discrepancies through the framework of Markov decision processes (MDPs).

We propose the Transferred Fitted Q-Iteration algorithm with general function approximation, which enables direct estimation of the optimal action-state function Q* using both target and source data. Under sieve approximation, we establish the relationship between statistical performance and MDP task discrepancy, highlighting the influence of source and target sample sizes and task discrepancy on the effectiveness of knowledge transfer. Our theoretical and empirical results demonstrate that the final learning error of the Q* function is significantly reduced compared to the single-task learning rate.

A transfer learning approach for interval-censored failure time data

♦(Tony) Jianguo Sun

University of Missouri

In this talk, we propose a transfer learning approach for

regression analysis of interval-censored failure time data. The proposed

approach is shown to be effective and applied to a motivating example.

Applications of Bayesian Power Prior with a Discount Function in Medical Device Trials

♦Hong Zhao

Abbott

In this study, we explored a Bayesian analysis to estimate the between-group hazard ratio using data from a previous trial. This approach helps us make better use of existing information to plan the current trial. A modified "power prior" model with a discount function was used to combine the old and new data, which is a novel method to adjust the influence of the old data.

Simulations were performed to determine the approximate sample size savings that could result from using this Bayesian analysis with an informative prior compared to the sample size required by a frequentist analysis for the same level of statistical power. Our findings showed that using this Bayesian method can significantly reduce the sample size without losing statistical power.

This approach can make clinical trials more efficient and cost-effective, ultimately speeding up the process of bringing new medical treatments to patients.

Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Survival Outcome for Large Data

Shanpeng Li, Emily Ouyang, Jin Zhou, Xinping Cui, ♦Gang Li

City of Hope, UCR, University of California at Los Angeles, UCR, University of California at Los Angeles

Despite the explosive growth of literature on joint models to correlate longitudinal and time-to-event data, efficient implementation of jointly modeling multiple biomarkers and time-to-event outcome has lagged behind, and their current implementations do not scale to large datasets with tens of thousands to millions of subjects. To address this, we propose a fast approximate expectation-maximization (EM) algorithm for a semiparametric joint model that handles multiple biomarkers and competing risks time-to-event outcome. The fast approximate EM algorithm utilizes both customized linear scan algorithms and a normal approximation of the posterior distribution of random effects, significantly reducing the computational burdens by a factor of up to hundreds of thousands compared to the existing approaches, often reducing the runtime from days to minutes. We validate the accuracy and efficiency of our approximation method through various simulation studies and further demonstrate its practical applications by using a real world large-scale Biobank study.

25CHI095: Statistical Advances for Integrative Multi-Omics Data Analysis

Local genetic correlation via knockoffs reduces confounding due to cross-trait assortative mating

♦Shiyang Ma, Fan Wang, Iuliana Ionita-Laza

Shanghai Jiao Tong University, Columbia University, Columbia University

Local genetic correlation analysis is an important tool for identifying genetic loci with shared biology across traits. Recently Border et al. have shown that the results of these analyses are confounded by cross-trait assortative mating (xAM) leading to many false positive findings. Here we describe LAVA-Knock, a local genetic correlation method that builds off an existing genetic correlation method, LAVA, and augments it by generating synthetic data in a way that preserves local and long-range linkage disequilibrium (LD), allowing us to reduce the confounding induced by xAM. We show in simulations based on a realistic xAM model and in GWAS applications for 630 trait pairs that LAVA-Knock can greatly reduce the bias due to xAM relative to LAVA. Furthermore, we show a significant positive correlation between the reduction in local genetic correlations and estimates in the literature of cross-mate phenotype correlations; in particular, pairs of traits that are known to have high cross-mate phenotype correlation values have a significantly higher reduction in the number of local genetic correlations compared with other trait pairs. A few representative examples include education and intelligence, education and alcohol consumption, and attention-deficit hyperactivity disorder and depression. These results suggest that LAVA-Knock can reduce confounding due to both short range LD but also long-range LD induced by xAM.

scHiCPRSiM: single-cell Hi-C Practical and Rational SiMulator

♦Huiling Liu, Rui Ma, Xingping Cui, Wenxiu Ma

South China University of Technology

Recent advancements in single-cell Hi-C (scHi-C) techniques have empowered us to explore the three-dimensional genome organization at the individual cell level. Various computational and statistical methods have been developed for scHi-C data analysis; However, benchmarking these methods remains a significant challenge, primarily due to the scarcity of high-quality scHi-C datasets.

To address this challenge, we propose scHiCPRSiM, a versatile and robust statistical simulator of scHi-C data. scHiCPRSiM serves as a simulation-based framework that enables researchers to quantitatively assess scHi-C experimental design and benchmark existing analytical approaches. Notably, scHiCPRSiM excels in generating realistic scHi-C datasets that closely resemble real data, capturing vital chromatin structure features, providing valuable guidance for optimizing experimental design by striking a balance between cell clustering accuracy and budget constraints, and facilitating the performance evaluation and comparison of scHi-C analytical methods.

Spatial Resolved Gene Regulatory Networks Analysis

Ishita Debnath, ♦Zhana Duren

Indiana University, Indiana University

Integrating spatial transcriptomics – which maps gene expression location within tissues – with single-cell multi-omics data, profiling gene expression and chromatin accessibility (or other epigenomic data), offers powerful insights into gene regulation. However, commercially available kits for simultaneous spatial multi-omics profiling are currently unavailable, hindering widespread data generation. Here, we present ISON (Integrated Spatial Omics Networks), a novel computational method to infer spatial-resolved gene regulatory networks by leveraging existing single-cell multiome data and spatial transcriptomics data. ISON accurately predicts omics profiles for spatial spots and reconstructs spatially resolved gene regulatory networks, demonstrating scalability in both time and memory. Importantly, ISON omics prediction preserves cis- and trans- regulatory information and enables estimation of transcription factor (TF) activity at the spot level, distinguishing between TFs even within the same family – a capability absents in approaches relying solely on ATAC-seq data. Application of ISON to Alzheimer's disease data reveals disease- and age-specific spatial gene regulatory modules, highlighting its potential for uncovering spatially organized mechanisms driving complex biological processes.

Clustering-free nuanced marker identification and attribution and its application in the taurine compensatory effect discovery in dilated cardiomyopathy

Jinpu Cai, Xiaorui Liu, Cheng Wang, Yibo Zhang, Luting Zhou, Ziqi Rong, Hongbin Shen, Qiuyu Lian, Liang Chen, ♦Hongyi Xin

Shanghai Jiao Tong University, Fuwai Hospital, Shanghai Jiao Tong University, Fuwai Hospital, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Cambridge University, Fuwai Hospital, Shanghai Jiao Tong University

Advances in single-cell sequencing technology have facilitated the revelation of molecular heterogeneity at cellular resolution. However, identifying nuanced pathological subpopulations and their markers remains challenging due to high-dimensional data complexity, clustering-dependent workflows, and the entanglement of major- and sub-type markers in conventional differential expression analyses. These limitations obscure critical yet subtle biologically variations, such as pathological-specific cell states.

Here, we present TAMER (Taxonomy Aware Marker Extraction and Registration), a clustering-free algorithm that identifies and organizes cell-type markers by decoding the spectra of inter-gene mutual exclusivity—a robust statistic for detecting cell-type markers at all granularities. TAMER identifies and categorizes putative markers into modules and organizes modules into a putative major-to-sub-type hierarchy, while completely circumventing clustering biases. With the marker hierarchy, TAMER can guide transparent and sensible top-down hierarchical clustering that partitions cells into progressively finer-grained sub-populations; enables nuanced-biological-variation-preserving batch integration; and supports cell type distribution highlighting in spatial transcriptomics.

We applied TAMER to human dilated cardiomyopathy (DCM) snRNA-seq data and identified two new pathological cardiomyocyte sub-population markers, TauT (taurine transporter), and Spock1. We hypothesized that TauT upregulation as a stress-coping self-rescue mechanism of cardiomyocytes under excessive oxidative stress. Taurine supplementation in cardiomyocyte cell-line alleviated the pathological phenotype, which is reversed by subsequent TauT knockdown. In a mouse model, taurine supplementation delayed pathological ventricular remodeling and improved survival. Our findings established TAMER as a paradigm-shifting tool for pathological subpopulation marker discovery and underscore the therapeutic potential of taurine supplementation in DCM treatment.

25CHI102: Statistical learning based on high dimensional and complex data

Statistical Approaches to MLP Approximation in Efficient Language Models

♦Yifan Chen

Hong Kong Baptist University

Transformer is the backbone architecture of most recent phenomenal language models. In this talk, I will delve into the approximation techniques for the MLP modules in Transformers. Firstly, I will discuss the compression of the MLP layers in Transformers, which preserves the neural tangent kernel (NTK) thereof and accelerates both fine-tuning and inference for large language models. Next, for new models similar to DeepSeek-V3, their Mixture-of-Experts modules require tremendous GPU memory; I will reveal the distributional representation of expert modules in Mixture-of-Experts models and thus apply distributional techniques to model the parameter compression process. The two aspects collectively showcase the statistical structures behind popular deep learning designs.

Local Information for Global Network Estimation in Latent Space Models

♦Lijia Wang, Xiao Han, Yanhui Wu, Y. X. Rachel Wang

City University of Hong Kong, University of Science and Technology of China, University of Hong Kong, University of Sydney

In social networks, understanding an individual's local neighborhood is key to analyzing their behavior. This paper explores using a partial information network centered on an individual to estimate the global network by fitting a general latent space model. The partial network often lacks many edges due to a random, sparse neighborhood, presenting significant estimation challenges. To address this, we propose a projected gradient descent algorithm to maximize the observed data's likelihood and provide theoretical convergence guarantees under various neighborhood structures. Our findings reveal how bias in an individual's local perspective affects estimation, quantified by an imbalance measure. Simulated and real network tests demonstrate our method's effectiveness and offer insights into network structures, such as the tradeoff between degrees and imbalance.

Statistical inference for functional data over multi-dimensional domain

Qirui Hu, ♦Lijian Yang

Tsinghua University, Tsinghua University

This work develops inference tools for the mean function of functional data over a multi-dimensional domain. A two-step mean estimator based on tensor product spline estimates of individual trajectories is shown oracally efficient, i.e., it is asymptotically indistinguishable from the infeasible estimator using unobservable trajectories. Consistent estimates of covariance function as well as exact quantile of the limiting maximal deviation are obtained by innovative use of results on sharp comparison of Gaussian extreme distributions and quantiles, leading to asymptotic coverage and order $n^{-1/2}$ uniformly adaptive width of data-driven simultaneous confidence regions (SCRs). Also formulated are one-sided SCRs that can be used for testing against uniform upper and lower bound of the mean function. Extensive Monte Carlo experiments corroborate the theory, and a satellite ocean dataset collected by Copernicus Marine Environment Monitoring Service (CMEMS) illustrates how the proposed SCR is used.

25CHI109: Statistical Network Analysis and Applications

Modelling Homophily in Autoregressive Networks

♦*Xinyang Yu*

London School of Economics and Political Science

Statistical modeling of network data is an important topic in various areas. Although many real networks are dynamic in nature, most existing statistical models and related inferences for network data are confined to static networks, and the development of the foundation for dynamic network models is still in its infancy. In particular, to the best of our knowledge, no attempts have been made to jointly address node heterogeneity and link homophily among dynamic networks. Being able to capture these network features simultaneously will only bring new insights on understanding how networks were formed, but also provide more sophisticated tools for the prediction of a future network with statistical guarantees. In this project, we adopt an autoregressive formulation for dynamic networks, which specifically depicts the dynamic change of the edges over time with joint consideration of node heterogeneity and link homophiy. In particular, our model accounts for link homophily associated with both observed traits and latent traits of the nodes. A novel convex loss based framework is constructed to generate stable estimations for the high dimensional parameters. As a byproduct, the estimated latent traits and the latent traits are further used for community detection when there is a stochastic block structure under the networks.

A network approach to compute hypervolume under receiver operating characteristic manifold for multi-class biomarkers

♦*Qunqiang Feng, Pan Liu, Pei-Fen Kuan, Fei Zou, Jianan Chen, Jialiang Li*

University of Science and Technology of China, National University of Singapore, Stony Brook University, University of North Carolina at Chapel Hill, National University of Singapore, National University of Singapore

Computation of hypervolume under ROC manifold (HUM) is necessary to evaluate biomarkers for their capability to discriminate among multiple disease types or diagnostic groups. However the original definition of HUM involves multiple integration and thus a medical investigation for multi-class receiver operating characteristic (ROC) analysis could suffer from huge computational cost when the formula is implemented naively. We introduce a novel graph-based approach to compute HUM efficiently in this article. The computational method avoids the time-consuming multiple summation when sample size or the number of categories is large. We conduct extensive simulation studies to demonstrate the improvement of our method over existing R packages. We apply our method to two real biomedical data sets to illustrate its application.

Community detection in weighted networks via the profile-pseudo likelihood method

Yang Liu, Jiangzhou Wang, ♦*Binghui Liu*

Northeast Normal University, Shenzhen University, Northeast Normal University

In this paper, we consider the issue of community detection in weighted networks. Most methods addressing this issue, particularly those statistical approaches based on likelihood optimization, suffer from a notable drawback: the necessity to specify in advance the particular form of the distribution of edge weights conditional on the community labels. This requirement dictates that algorithms based on likelihood optimization must be custom-tailored exclusively to the specific form of distribution, which exhibits significant limitations in practical applications where the form of distribution is unknown. To address this limitation, we propose two novel methods based on the expectation profile-pseudo likelihood maximization, for community detection in both undirected and directed weighted networks, which are applicable to various types of weighted networks and independent of the specific form of the conditional distribution of the edge weights. In theory, we establish weak and strong consistency, respectively, of the resulting community label estimations within the sub-exponential family, and then establish the convergence of the proposed algorithms. In simulation studies, we demonstrate significant advantages of the proposed methods across a wide range of conditional distributions and parameter settings, both in terms of community detection accuracy and computational efficiency. In practical applications, we showcase the applicability of the proposed methods on three real-world weighted networks.

Transfer Learning Under High-Dimensional Network Convolutional Regression Model

Liyuan Wang, Jiachen Chen, Kathryn Lunetta, ♦*Danyang Huang, Huimin Cheng, Debarghya Mukherjee*

Renmin University of China

Transfer learning enhances model performance by utilizing knowledge from related domains, particularly when labeled data is scarce. While existing research addresses transfer learning under various distribution shifts in independent settings, handling dependencies in networked data remains challenging. To address this challenge, we propose a high-dimensional transfer learning framework based on network convolutional regression (NCR), inspired by the success of graph convolutional networks (GCNs). The NCR model incorporates random network structure by allowing each node’s response to depend on its features and the aggregated features of its neighbors, capturing local dependencies effectively. Our methodology includes a two-step transfer learning algorithm that addresses domain shift between source and target networks, along with a source detection mechanism to identify informative domains. Theoretically, we analyze the lasso estimator in the context of a random graph based on the Erdős–Rényi model assumption, demonstrating that transfer learning improves convergence rates when informative sources are present. Empirical evaluations, including simulations and a real-world application using Sina Weibo datademonstrate substantial improvements in prediction accuracy, particularly when labeled data in the target domain is limited.

25CHI110: Statistical theory of neural networks

Rates for least squares using over-parameterized neural networks

♦*Yunfei Yang*

Sun Yat-sen University

Recent studies showed that deep neural networks can achieve minimax optimal rates for learning smooth function classes. However, most of these results require that the neural networks in use are under-parameterized, which cannot explain the successes of over-parameterized models used in practice. In this talk, we will discuss how to derive convergence rates for neural networks in the over-parameterized regime. We will begin with a discussion on the approximation capacity of ReLU neural networks with certain norm constraints on the weights. By using this result, we are able to prove nearly optimal learning rates for least squares estimations based on over-parameterized (deep or shallow) neural networks if the weights are properly constrained. Finally, we will also show how to obtain minimax optimal rates for shallow neural networks by using localization technique and generalize the results to regularized least squares.

Solving PDEs on Spheres with Physics-Informed Convolutional Neural Networks

♦*Lei Shi*

Fudan University

Physics-informed neural networks (PINNs) have been demonstrated to be efficient in solving partial differential equations (PDEs) from a variety of experimental perspectives. Some recent studies have also proposed PINN algorithms for PDEs on surfaces, including spheres. However, theoretical understanding of the numerical performance of PINNs, especially PINNs on surfaces or manifolds, is still lacking. In this talk, we establish a rigorous analysis of the physics-informed convolutional neural network (PICNN) for solving PDEs on the sphere. By using and improving the latest approximation results of deep convolutional neural networks and spherical harmonic analysis, we prove an upper bound for the approximation error with respect to the Sobolev norm. Subsequently, we integrate this with innovative localization complexity analysis to establish fast convergence rates for PICNN. Our theoretical results are also confirmed and supplemented by our experiments. In light of these findings, we explore potential strategies for circumventing the curse of dimensionality that arises when solving high-dimensional PDEs.

Optimization and Generalization of Gradient Methods for Shallow Neural Networks

♦*Yunwen Lei, Yiming Ying, Ding-Xuan Zhou*

The University of Hong Kong, University of Sydney, University of Sydney

Neural networks have achieved impressive performance in various applications. In this talk, we discuss the optimization and generalization of shallow neural networks (SNNs). We consider both gradient descent (GD) and stochastic gradient descent (SGD) to train SNNs. We show how the optimization and generalization should be balanced to obtain consistent error bounds under a relaxed overparameterization setting.

Nonparametric GARCH: A Deep Learning Approach

Ruizhi Deng, ♦*Guohao Shen, Ngai Hang Chan*

The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, The City University of Hong Kong

This paper introduces a novel approach to estimating nonparametric GARCH models using deep neural networks. We propose an efficient iterative algorithm for training these deep estimators, characterized by ease of implementation and adaptability to various model settings and loss functions. We establish learning guarantees for the proposed method, including non-asymptotic upper bounds on prediction error under mild assumptions. Notably, we demonstrate that our deep neural network estimator can adapt to the true lag dimension of the volatility model even when the input dimension is overspecified. This crucial property ensures optimal performance even with suboptimal input choices. We validate the effectiveness of our approach through extensive simulations, showcasing its superiority over competing methods, particularly in high-dimensional, nonlinear, and complex volatility scenarios. We further demonstrate the practical utility of our deep nonparametric GARCH estimator by applying it to real-world financial data.

25CHI113: Theoretical Advances in Machine Learning and Dimension Reduction, and Functional Data Analysis

On the structural dimension of sliced inverse regression

♦*Dongming Huang, Songtao Tian, Qian Lin*

National University, Tsinghua University, Tsinghua University

The central space of a joint distribution (X, Y) is the minimal subspaceS such that Y and X are conditionally independent given PX, where P is the projection onto S. Sliced inverse regression (SIR), one of the most popular methods for estimating the central space, often performs poorly when the structural dimension d is large. In this paper, we demonstrate that the generalized signal-noise-ratio (gSNR) tends to be extremely small for a general multiple-index model when d is large. Then we determine the minimax rate for estimating the central space over a large class of high dimensional distributions with a large structural dimension d (i.e., there is no constant upper bound on d) in the low gSNR regime. This result not only extends the existing minimax rate results for estimating the central space of distributions with fixed d to that with a large d, but also clarifies that the degradation in SIR performance is caused by the decay of signal strength. The technical tools developed here might be of independent interest for studying other central space estimation methods.

Diagonal Over-parameterization in Reproducing Kernel Hilbert Spaces as an Adaptive Feature Model: Generalization and Adaptivity

♦*Yicheng Li, Qian Lin*

Department of Statistics and Data Science, Tsinghua University, Department of Statistics and Data Science, Tsinghua University

This paper introduces a diagonal adaptive kernel model that dynamically learns kernel eigenvalues and output coefficients simultaneously during training.

Unlike fixed-kernel methods tied to the neural tangent kernel theory, the diagonal adaptive kernel model adapts to the structure of the truth function, significantly improving generalization over fixed-kernel methods, especially when the initial kernel is misaligned with the target.

Moreover, we show that the adaptivity comes from learning the right eigenvalues during training, showing a feature learning behavior.

By extending to deeper parameterization, we further show how extra depth enhances adaptability and generalization.

This study combines the insights from feature learning and implicit regularization and provides new perspective into the adaptivity and generalization potential of neural networks beyond the kernel regime.

Duality Between Context Data and Model Parameters in Transformers

Brian Chen, Hui Jin, Haonan Wang, Kenji Kawaguchi, ♦Tianyang Hu

National University of Singapore, Huawei Noah's Ark Lab, National University of Singapore, National University of Singapore, National University of Singapore

In-Context Learning (ICL) has emerged as a powerful capability of Large Language Models (LLMs), presenting intriguing theoretical questions in the machine learning community. This talk explores fundamental relationships between ICL and model weight updates in transformer-based LLMs. We first analyze linear transformers, demonstrating that ICL functions as a greedy layer-wise gradient descent algorithm in parameter space. Building on this, we show that while exact internalization of ICL is impossible within standard architectures, it can be achieved through the introduction of novel query-dependent bias terms in attention modules. For standard transformers, we extend these insights and present methods for approximate conversions via attention linearization in a training-free fashion, or more generally, through a specialized attention memory module that can be trained via supervised-finetuning. These findings reveal deeper connections between contextual information and model parameters, suggesting new algorithmic approaches for more effective model adaptation.

Modified Tests of Linear Hypotheses Under Heteroscedasticity for Multivariate Functional Data with Finite Sample Sizes

♦Tianming Zhu

National Institute of Education, Nanyang Technological University

As big data continues to grow, statistical inference for multivariate functional data (MFD) has become crucial. Although recent advancements have been made in testing the equality of mean functions, research on testing linear hypotheses for mean functions remains limited. Current methods primarily consist of permutation-based tests or asymptotic tests. However, permutation-based tests are known to be time-consuming, while asymptotic tests typically require larger sample sizes to maintain an accurate Type I error rate. This paper introduces three finite-sample tests that modify traditional MANOVA methods to tackle the general linear hypothesis testing problem for MFD. The test statistics rely on two symmetric, nonnegative-definite matricesapproximated by Wishart distributions, with degrees of freedom estimated via a U-statistics-based method. The proposed tests are affine-invariant, computationally more efficient than permutation-based tests, and better at controlling significance levels in small samples compared to asymptotic tests. A real-data example further showcases their practical utility.

25CHI037: Innovative Inference Methods for Complex Data: Bridging Theory and Practice

A practical interval estimation method for spectral density function

♦Haihan Yu, Mark Kaiser, Daniel Nordman

University of Rhode Island, Iowa State University, Iowa State University

The spectral density function can play a key role in time series analysis, where nonparametric interval estimation of the spectral density is a fundamental issue. However, the prevailing pointwise interval methods for spectral densities, including chi-square approximation and frequency domain bootstrap (FDB), can be misleading in practice, perhaps more so than appreciated, as confidence intervals often exhibit low coverage accuracy as well as high sensitivity to tuning parameters. To provide a practical alternative, we propose a new hybrid method that combines the strengths of empirical likelihood (EL) and FDB. The method involves developing an EL statistic for spectral density inference along with a corresponding bootstrap approximation under time dependence, where we allow for general time processes as well as for two different types of kernel smoothing found in application (so-called A- or K-windows). Such windows require differing theories and implementations in practice. As an advantage, the FDB-EL procedure is formally valid under mild conditions for application to a broad range of processes, including both linear and nonlinear time series. Simulation studies demonstrate that FDB-EL-based confidence intervals are effective compared to other methods, as intervals maintain good coverage accuracy while being less sensitive to bandwidth parameters. The confidence interval procedure is illustrated with an application to studying the wind spectrum. Extension to simultaneous confidence intervals has also been discussed.

A Semiparametric Causal Estimator without Ignorability

♦Guoliang Ma, Cindy Yu, Zhonglei Wang

Xiamen University, Iowa State University, Xiamen University

For observational studies, the assignment to treatment is often affected by auxiliary information or even the unobserved outcome of interest. Existing works for causal inference mainly assume ignorable assignment, and it may lead to erroneous inference if such an assumption is wrong. In this paper, we propose a novel semiparametric iterative procedure for causal inference based on a general non-ignorable assignment assumption, and it (a) can consistently estimate the model parameters with controllable bias, (b) allows for heterogeneity among individuals, and (c) is flexible in identifying a wide range of causal effects. Limiting properties of the proposed procedure are rigorously investigated. We conduct extensive simulation studies under various settings and demonstrate the efficacy of the proposed ES algorithm. Application to the National Longitudinal Survey of the Young Men on the 1997 cohort reveals that the 1997 cohort relied on potential outcomes to make college-entrance decisions. On average, the return on college experience is large and grows over the years.

Online Tensor Inference

♦Xin Wen, Will Wei Sun, Yichen Zhang

New York University, Purdue University, Purdue University

Contemporary applications, such as recommendation systems and mobile health monitoring, require real-time processing and analysis of sequentially arriving high-dimensional tensor data. Traditional offline learning, involving the storage and utilization of all data in each computational iteration, becomes impractical for these tasks. Furthermore, existing low-rank tensor methods lack the capability for online statistical inference, which is essential for real-time predictions and informed decision-making. This paper addresses these challenges by introducing a novel online inference framework for low-rank tensors. Our approach employs Stochastic Gradient Descent (SGD) to enable efficient real-time data processing without extensive memory requirements. We establish a non-asymptotic convergence result for the online low-rank SGD estimator, nearly matches the minimax optimal estimation error rate of offline models. Furthermore, we propose a simple yet powerful online debiasing approach for sequential statistical inference. The entire online procedure, covering both estimation and inference, eliminates the need for data splitting or storing historical data, making it suitable for on-the-fly hypothesis testing. In our analysis, we control the sum of constructed super-martingales to ensure estimates along the entire solution path remain within the benign region. Additionally, a novel spectral representation tool is employed to address statistical dependencies among iterative estimates, establishing the desired asymptotic normality.

25CHI048: Modern Machine Learning: Tackling Real-World Data Challenges

Renewable l1-regularized linear support vector machine with high-dimensional streaming data

Na Zhang, Jinhan Xie, Xiaodong Yan, Bei Jiang, Ting Li, ♦Linglong Kong

University of Alberta, Yunan University, Xi'an Jiaotong University, University of Alberta, Hong Kong Polytechnic University, University of Alberta

The rapid growth of modern data collection methods brings new challenges for existing classification problems and the storage of huge datasets in memory. The need to develop online update methods is becoming increasingly pressing. In this paper, we study the renewable estimation process for a linear support vector machine (SVM) in high-dimensional online settings. The proposed renewable estimation process, which includes online ℓ_1-regularized and online debiased procedures, is feasible for handling high-dimensional streaming data since the online estimators are updated by integrating current new data batches with summary statistics of historical data, rather than re-accessing the entire raw dataset. Theoretically, we prove the convergence rates of the proposed online estimators under mild conditions. Numerical studies confirm the effectiveness of the proposed methods.

Inference of Comparing Generative AI Models

♦Zijun Gao, Yan Sun

University of Southern California, University of Pennsylvania

Generative AI (GenAI) models have recently achieved remarkable empirical performance in various applications, however, their evaluations yet lack uncertainty quantification. In this paper, we propose a method to compare two generative models based on an unbiased estimator of their relative performance gap. Statistically, our estimator achieves parametric convergence rate and asymptotic normality, which enables valid inference. Computationally, our method is efficient and can be accelerated by parallel computing and leveraging pre-storing intermediate results. On simulated datasets with known ground truth, we show our approach effectively controls type I error and achieves power comparable with commonly used metrics. Furthermore, we demonstrate the performance of our method in evaluating diffusion models on real image datasets with statistical confidence.

Transfer Reinforcement Learning: Value-Based Methods for Non-Stationary MDPs

♦Elynn Chen

New York University

In dynamic decision-making scenarios across business, healthcare, and education, leveraging data from diverse populations can significantly enhance reinforcement learning (RL) performance for specific target populations, especially when target samples are limited. We develop comprehensive frameworks for transfer learning in RL, addressing both stationary Markov decision processes (MDPs) with iterative Q*-learning and non-stationary finite-horizon MDPs with backward inductive Q*-learning.

For stationary MDPs, we propose an iterative Q*-learning algorithm with knowledge transfer, establishing theoretical justifications through faster convergence rates under similarity assumptions. For non-stationary finite-horizon MDPs, we introduce two key innovations: (1) a novel "re-weighted targeting procedure" that enables cross-satege transfer along multiple temporal steps, and (2) transfer deep Q*-learning that leverages neural networks as function approximators. We demonstrate that while naive sample pooling strategies may succeed in regression settings, they fail in MDPs, necessitating our more sophisticated approach. We establish theoretical guarantees for both settings, revealing the relationship between statistical performance and MDP task discrepancy. Our analysis illuminates how source and target sample sizes impact transfer effectiveness. The framework accommodates both transferable and non-transferable transition density ratios while assuming reward function transferability. Our analytical techniques have broader implications, extending to supervised transfer learning with neural networks and domain shift scenarios. Empirical evidence from both synthetic and real datasets validates our theoretical results, demonstrating significant improvements over single-task learning rates and highlighting the practical value of strategically constructed transferable RL samples in both stationary and non-stationary contexts.

This talk is based on the following two papers.

Transfer Q-Learning: https://arxiv.org/abs/2202.04709

Deep Transfer -Learning for Offline Non-Stationary Reinforcement Learning: https://arxiv.org/abs/2501.04870

Data-Driven Knowledge Transfer in Batch Q* Learning: https://arxiv.org/abs/2404.15209

Transition Transfer -Learning for Composite Markov Decision Processes: https://arxiv.org/abs/2502.00534

Departments of Data Science and AI, and Applied Mathematics

♦Jian Huang

The Hong Kong Polytechnic University

Continuous Normalizing Flows (CNFs) are a generative modeling technique that utilizes ordinary differential equations to learn probability distributions. This approach has been successful in a range of applications, including image synthesis, protein structure prediction, and molecule generation. In this talk, we will present the CNF method and explore its theoretical properties through a flow matching objective function. We will then introduce a conditional CNF method and demonstrate its application in controlled image generation by fine-tuning Stable Diffusion 3, a large foundational image model.

25CHI049: Modern Multivariate Analysis for Tensor and Multiview Data

D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High-dimensional Data

♦Hai Shu, Hongtu Zhu

New York University, The University of North Carolina at Chapel Hill

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of multiple high-dimensional data views is to decompose each view's data matrix into three parts: a low-rank common-source matrix generated by common latent factors of all data views, a low-rank distinctive-source matrix generated by distinctive latent factors of the corresponding data view, and an additive noise matrix. Existing decomposition methods often focus on the uncorrelatedness between the common latent factors and distinctive latent factors, but inadequately address the equally necessary uncorrelatedness between distinctive latent factors from different data views. We propose a novel decomposition method, called Decomposition of Common and Distinctive Latent Factors (D-CDLF), to effectively achieve both types of uncorrelatedness. Consistent estimators of our D-CDLF method are established with good finite-sample numerical performance. The superiority of D-CDLF over state-of-the-art methods is also corroborated in simulations and real-world data analysis.

Low-rank decomposition, dimension-reduction subspaces and tensor change detection

Jiaqi Huang, ♦Ning Wang, Lixing Zhu

Beijing Normal University

We introduce a novel two-stage approach for analyzing tensor sequences exhibiting structural changes, leveraging Tucker decomposition to simultaneously identify low-rank and sparse structures. Our method first transforms tensors into a rank-based sequence to detect changes in projection matrices, employing an eigen-decomposition-based estimation that has good performance in accuracy and consistency. Subsequently, we address core tensor changes, developing a "Completion Algorithm" to handle varying sizes in the sequence and improve detection robustness. By precisely categorizing change points and mitigating redundancy, we achieve computationally efficient detection. Theoretical analysis and empirical validation demonstrate the method's efficacy in analyzing dynamic high-order tensor data, with further extensions to handle structural heterogeneity.

Statistical Inference for Low-Rank Tensor Models

♦Ke Xu, Elynn Chen, Yuefeng Han

University of Notre Dame, New York University, University of Notre Dame

Statistical inference for tensors has emerged as a critical challenge in analyzing high-dimensional data in modern data science. This paper introduces a unified framework for inferring general and low-Tucker-rank linear functionals of low-Tucker-rank signal tensors for several low-rank tensor models. Our methodology tackles two primary goals: achieving asymptotic normality and constructing minimax-optimal confidence intervals. By leveraging a debiasing strategy and projecting onto the tangent space of the low-Tucker-rank manifold, we enable inference for general and structured linear functionals, extending far beyond the scope of traditional entrywise inference. Specifically, in the low-Tucker-rank tensor regression or PCA model, we establish the computational and statistical efficiency of our approach, achieving near-optimal sample size requirements (in regression model) and signal-to-noise ratio (SNR) conditions (in PCA model) for general linear functionals without requiring sparsity in the loading tensor. Our framework also attains both computationally and statistically optimal sample size and SNR thresholds for low-Tucker-rank linear functionals. Numerical experiments validate our theoretical results, showcasing the framework's utility in diverse applications. This work addresses significant methodological gaps in statistical inference, advancing tensor analysis for complex and high-dimensional data environments.

Sparse and integrative principal component analysis for multiview data

Lin Xiao, ♦Luo Xiao

North Carolina State University, North Carolina State University

We consider dimension reduction of multiview data, which are emerging in scientific studies. Formulating multiview data as multivariate data with block structures corresponding to the different views, or views of data, we estimate top eigenvectors from multiview data that have two-fold sparsity, elementwise sparsity and blockwise sparsity. We propose a Fantope-based optimization criterion with multiple penalties to enforce the desired sparsity patterns and a denoising step is employed to handle potential presence of heteroskedastic noise across different data views. An alternating direction method of multipliers (ADMM) algorithm is used for optimization. We derive the l2 convergence of the estimated top eigenvectors and establish their sparsity and support recovery properties. Numerical studies are used to illustrate the proposed method.

25CHI013: Advances in Statistical Learning and Network Analysis for Complex Data

Adaptive Block-Based Change-Point Detection for Sparse Spatially Clustered Data with Applications in Remote Sensing Imaging

Alan Moore, Lynna Chu, ♦Zhengyuan Zhu

Iowa State University, Iowa State University, Iowa State University

We present a non-parametric change-point detection approach to detect potentially sparse changes in a time series of high-dimensional observations or non-Euclidean data objects. We target a change in distribution that occurs in a small, unknown subset of dimensions, where these dimensions may be spatially correlated. Our work is motivated by a remote sensing application where changes occur in small, spatially clustered regions over time. An adaptive block-based change-point detection framework is proposed that accounts for spatial dependencies across dimensions and leverages these dependencies to boost detection power and improve estimation accuracy. Through simulation studies, we demonstrate that our approach has superior performance in detecting sparse changes in datasets with spatial or local group structures. An application of the proposed method to detect changes in remote sensing imagery is presented.

Information-incorporated Network Construction with FDR Control

Hao Wang, Yumou Qiu, ♦Peng Liu

Iowa State University, Peking University, Iowa State University

Large-scale gene expression studies allow gene network construction to uncover associations among genes. To study direct associations among genes, networks based on partial correlations are preferred over those on marginal correlations. However, FDR control for partial correlation-based network construction is not well-studied. In addition, currently available partial correlation-based methods cannot take existing biological knowledge to help network construction while controlling FDR. In this talk, we propose a method called Partial Correlation Graph with Information Incorporation (PCGII). PCGII estimates partial correlations between each pair of genes by regularized node-wise regression that can incorporate prior knowledge while controlling the effects of all other genes. It handles high-dimensional data where the number of genes can be much larger than the sample size and controls FDR at the same time. We compare PCGII with several existing approaches through extensive simulation studies and demonstrate that PCGII has better FDR control and higher power. We apply PCGII to a plant gene expression dataset where it recovers confirmed regulatory relationships and a hub node, as well as several direct associations that shed light on potential functional relationships in the system. We also introduce a method to supplement observed data with a pseudogene to apply PCGII when no prior information is available, which also allows checking FDR control and power for real data analysis.

Bias-corrected Byzantine-robust Estimator via Cornish-Fisher Expansion for Distributed Learning

Zhixiang Zhou, Yibo Yuan, Xiaojun Mao, ♦Zhonglei Wang

Xiamen University, Xiamen University, Shanghai Jiao Tong University, Xiamen University

For distributed learning, median-of-means estimators are popular for statistical inference against Byzantine attacks, but they suffer from inefficiency or even bias when the sample is asymmetrically distributed. We tackle this problem by proposing a bias-corrected Byzantine-robust estimator via Cornish-Fisher expansion, and it can be widely implemented in various gradient descent algorithms for state-of-the-art deep learning models. We rigorously demonstrate that the bias of the proposed estimator is much smaller than a vanilla median-of-means estimator and its variations under regularity conditions, and asymptotic properties of the proposed estimator are also established. The proposed estimator outperforms its alternatives numerically in terms of bias and variance under different synthetic setups, and performs the best on both classification and regression tasks when analyzing the Taiwan Bankruptcy dataset and Communities and Crime dataset respectively.

Decentralized federated learning with fused lasso under distribution shift

Weidong Liu, Xiaojun Mao, ♦Xiaofei Zhang, Xin Zhang

Shanghai Jiao Tong University, Shanghai Jiao Tong University, Zhongnan University of Economics and Law, Meta Federated learning has gained increasing attention for enabling collaborative modeling with decentralized data. However, distribution shifts across local clients pose significant challenges to model robustness and generalization. To address this, we propose a decentralized federated learning framework with fused lasso regularization, which promotes both sparsity and smoothness among client-specific models. Our method explicitly accounts for local heterogeneity while leveraging shared structures across clients. We develop an efficient decentralized optimization algorithm for solving the fused lasso-regularized objective without relying on a central server. Theoretically, we establish the convergence rate and statistical consistency of the proposed estimator. Extensive simulation studies and real-world data applications validate the effectiveness and robustness of DecFL-FLasso under distribution shifts.

25CHI020: Advancing Risk Management with Statistical Learning

Exploratory Investment-Consumption with Non-Exponential Discounting

Yuling Chen, ♦Bin Li, David Saunders

University of Waterloo, University of Waterloo, University of Waterloo

We extend the classic Merton optimal investment-consumption problem to the reinforcement learning (RL) framework. Additionally, we incorporate a general non-exponential discounting function to capture the individual's risk preferences, which leads to time inconsistency in the exploratory control problem. Under standard entropy regularization and logarithmic utility, we obtain closed-form equilibrium investment-consumption policies. Specifically, the optimal investment policy follows a Gaussian distribution, while the optimal consumption policy follows a Gamma distribution. To validate our theoretical results, we develop and implement two RL algorithms—one based on the policy evaluation approach and the other on the q-learning approach—demonstrating their effectiveness through simulation studies.

Optimal Pooling of Catastrophe Risks

Minh Chau Nguyen, Tony Wirjanto, ♦Fan Yang

University of Waterloo, University of Waterloo, University of Waterloo

Catastrophe risk has long been recognized to pose a great threat to the insurance sector. Although natural disasters are rare events, they generally lead to devastating damages that traditional insurance schemes may not be able to efficiently cover. Catastrophe risk pooling is an effective way to diversify such risks that has recently been explored by multiple schemes such as the Caribbean Catastrophe Risk Insurance Facility. In this study, we explore a catastrophe risk pooling framework and study a set of approximations to optimal pooling strategies which allow all pool members to benefit from contributing their layer loss and sharing the aggregated risk together.

Defense Against Syntactic Textual Backdoor Attacks with Token Substitution

Xianwen He, Xinglin Li, Minhao Cheng, ♦Yao Li

University of North Carolina at Chapel Hill, University of North Carolina at Chapel Hill, Pennsylvania State University, University of North Carolina at Chapel Hill

Textual backdoor attacks present a substantial security risk to Large Language Models (LLM). It embeds carefully chosen triggers into a victim model at the training stage and makes the model erroneously predict inputs containing the same triggers as a certain class. Prior backdoor defense methods primarily target special token-based triggers, leaving syntax-based triggers insufficiently addressed. To fill this gap, this paper proposes a novel defense algorithm that effectively counters syntax-based as well as special token-based backdoor attacks. The algorithm replaces semantically meaningful words in sentences with entirely different ones but preserves the syntactic templates or special tokens, and then compares the predicted labels before and after the substitution to determine whether a sentence contains triggers. Experimental results confirm the algorithm's performance against these two types of triggers, offering a comprehensive defense strategy for model integrity.

The Joint Law of the Terminal Value, Running Maximum and Running Minimum of a Scalar Diffusion Process with Time-Inhomogeneous Drift

Philip Ernst, ♦Jixin Wang

Imperial College London, Imperial College London

The joint law of the running maximum, running minimum, and terminal value of a stochastic process plays a pivotal role in fields such as in mathematical finance for option pricing. To date, the research on the trivariate joint law of diffusions have been limited to the time-homogeneous case. In this paper, we describe our efforts to derive the trivariate joint law of (I_t, X_t, S_t) with a scalar diffusion and with time-inhomogeneous drift. Given mild regularity assumptions, we prove the existence of the three-dimensional joint density of (I_t, X_t, S_t). Moreover, we further derive the integral equations and the partial differential equations (PDEs) that the trivariate joint law should satisfy. Based on these equations, we demonstrate a sequence of L_1 approximations for the joint density. We also establish a numerical approximation scheme that performs well in two distinct real-world models.

25CHI032: Innovations in High Dimensional Complex Data Analysis: From Functional Data Analysis to Measurement Error Modeling

Addressing Misclassification in Outcome and Covariate via A Likelihood-Based Approach

Zhegn Yu, ♦Hua Shen

University of Calgary, University of Calgary

Misclassification of outcomes or categorical covariates often introduces serious bias into medical research, distorting relationships between exposures and outcomes. When both an outcome and a covariate are misclassified, these errors can severely undermine inference and obscure disease–exposure links. We present a likelihood-based algorithm for logistic regression that tackles dual misclassification by estimating sensitivity and specificity without requiring validation data. Extensive simulations under varied scenarios confirm its superiority over naive and ad hoc methods, and a real-world application highlights its practical value. This approach strengthens the reliability of inference in medical research and related fields.

High dimensional Recurrent Event Analysis with Error-Contaminated Covariates

♦Kaida Cai

Southeast University

In the analysis of recurrent event data with high-dimensional covariates, measurement error in covariates poses significant challenges for reliable variable selection and risk estimation. Naively ignoring measurement error often leads to biased and misleading results. To address this issue, we consider variable selection for recurrent event data with high-dimensional covariates contaminated by measurement error. We propose a penalized corrected likelihood approach to simultaneously adjust for measurement error and perform variable selection. Our method is based on a corrected score function that accounts for the measurement error through an additive error model, and a piecewise constant baseline hazard function. We establish the theoretical properties of the proposed estimator, including consistency and the oracle property. Simulation studies demonstrate that our method yields improved selection accuracy and reduced false discovery rate compared to naive approaches. We further illustrate the utility of our method using a real dataset on recurrent events.

Generalized SIMEX Method: Polynomial Approximation for Extrapolation

Li-Pang Chen, ♦Qihuang Zhang

National Chengchi University, McGill University

Measurement error is a common challenge in statistical analysis, often leading to incorrect parameter estimation. To address measurement error effects, the simulation and extrapolation (SIMEX) method is one of the widely used approaches because of its flexibility in model specification and generic scope of application. Key concerns of the SIMEX method include the number of repetitions in generating synthetic data and the choice of extrapolation function to recover the corrected estimates from the error-prone ones. In most of the existing developments, the quadratic function is frequently adopted as the extrapolation function. However, when measurement error effects are tremendously severe, quadratic functions may be suboptimal. In addition, the development of theoretical results of existing methods requires an unrealistic assumption that the true extrapolation function is known. To address those concerns, we propose GSIMEX, extending the SIMEX method by considering a higher-order polynomial function as the extrapolation function, which enables us to approximate the unknown and nonlinear extrapolation function. In addition, to improve the accuracy of the corrected estimator, we integrate subset selection and model averaging strategies. The theoretical results of GSIMEX, including the measurement of the approximation and asymptotic normality of the estimator, are rigorously established. Numerical studies are conducted for justification of validation, which show that GSIMEX is valid for dealing with severe measurement error effects and is flexible in handling different types of data structures and regression models. We analyze the simulated and spatial transcriptomics data to illustrate the usage of GSIMEX.

Meta-analyzing multiple functional data with functional fixed-effects model

Jiahao Tang, ♦Zongliang Hu, Hanbing Zhu, Yan Zhou, Shurong Zheng

Northeast Normal University, Shenzhen University, Anhui University, Shenzhen University, Northeast Normal University

Nowadays, many publicly available studies address similar research questions, making meta-analysis an efficient tool for comprehensively synthesizing information from these studies to achieve more reliable model estimation and interpretation. However, conducting a meta-analysis on multiple heterogeneous functional data sets poses significant challenges, particularly when the number of studies is small -- a common scenario in this context. In this paper, we propose a novel framework to address this challenge specifically for functional data. The proposed framework is based on the functional fixed-effects model, which extends the classical fixed-effects model to accommodate functional data settings. We also introduce a reparametrization procedure that decomposes the effect function in each study into a common-effect function and a study-specific deviation function. This approach facilitates the sharing of information across studies while simultaneously accounting for heterogeneity among them. To further enhance information borrowing and reduce noise in model estimation, we propose a variable selection procedure to identify null subregions of the common-effect and shrink study-specific deviation functions. Simulation studies and real data analysis demonstrate that our framework provides more reliable results and better model interpretation compared to existing meta-analysis methods.

25CHI035: Innovative Approaches in Electronic Health Record (EHR) Data Analysis

Age-Specific Outcome-guided Representation Learning for Patient Clustering with EHR Data

♦Linshanshan Wang, Mengyan Li, Molei Liu, Zongqi Xia, Tianxi Cai

Harvard University, Bentley University, Peking University, University of Pittsburgh, Harvard University

Stratifying patients into clinically meaningful subgroups using electronic health record (EHR) data is critical for advancing precision medicine, particularly in heterogeneous diseases such as Alzheimer's disease (AD). While prior work has applied unsupervised clustering to identify AD subtypes, these approaches often ignore outcome information during training and fail to account for age-related heterogeneity, a major axis of variation in disease manifestation and prognosis. To address these limitations, we propose SOLAR (age-Specific Outcome-guided representation Learning for pAtient clusteRing), a novel framework that integrates high-dimensional EHR features and survival outcomes while explicitly modeling age-group structure. SOLAR is built on a representation learning framework, under which we propose to jointly learn patient clusters across age groups, encouraging shared structure while allowing for age-specific flexibility. Through extensive simulations, we demonstrate the superior performance of our method in terms of classification, clustering and outcome prediction, compared with existing methods. We applied SOLAR to stratification of AD patients using EHR data, demonstrating that it is capable of identifying clinically meaningful clusters with distinct survival profiles and outperforms existing outcome-agnostic or age-unaware methods. Our results highlight the value of jointly modeling covariates and outcomes across age groups for robust and interpretable patient stratification.

An Evaluation Framework for Ambient Digital Scribing Tools in Clinical Applications

Haoyuan Wang, Rui Yang, Mahmoud Alwakeel, Ankit Kayastha, Anand Chowdhury, Joshua M. Biro, Anthony D. Sorrentino, Michael J. Pencina, Kathryn I. Pollak, ♦Chuan Hong

Duke University, Duke-NUS Medical School, Duke University, Duke University, Duke University, Medstar Health National Center for Human Factors in Healthcare, Duke University, Duke University, Duke University

Background. Ambient digital scribing (ADS) tools are transforming healthcare by reducing the burden of clinicians taking notes during clinical encounters, which might mitigate clinician burnout and turnover. Existing artificial intelligence (AI) -driven ADS tools automate transcription, diarization, and medical note generation, decreases documentation time and enhances clinical efficiency. Products like Abridge, Dax Copilot by Nuance, and Suki showcase ADS' potential across various specialties. As these AI-driven tools become integral to clinical workflows, robust digital governance frameworks are essential to ensure their ethical, secure, and effective deployment, preventing unsafe practices and ethical dilemmas in healthcare. In this study, we propose and test a comprehensive ADS evaluation framework incorporating human qualitative evaluation, automated quantitative metrics, and large language models (LLMs) as evaluators.

Methods. Our framework evaluates transcription, diarization, and medical note generation through multiple layers. Transcription and diarization accuracy were assessed based on human evaluation and automated evaluation. Medical notes were assessed across fluency, coherence, clarity, brevity, structuring, relevance, completeness, factuality, prudence based on a combination of human judgment, automated evaluation, and LLM-based evaluation. Additionally, simulation-based testing was conducted to ensure resilience against bias, fairness and adversarial challenges. To showcase the utility of the proposed evaluation framework, we internally developed a GPT-4o-based ADS tool and evaluated it using 40 audio recordings ranging from 6 to 35 minutes (average 16.7 minutes) from a clinical study on smoking cessation among pregnant patients.

Results. Quantitative results supported qualitative findings, demonstrating that the internally developed ADS tool demonstrated satisfactory overall performance. Fairness testing, which incorporated patient race, ethnicity, and social determinants of health (SDOH), showed no significant performance variations, in ROUGE, BLEU, or BERTScore across demographic subgroups, indicating unbiased and consistent outputs. However, toxicity scores (measuring the presence of harmful or offensive language in generated notes) were significantly lower for transcripts labeled “Black” versus “White” and “Black” versus “Asian” ($p < 0.0001$). Additionally, adversarial testing confirmed the tool's robustness in handling diarization errors, rare diseases, inappropriate content, and irrelevant content. However, it identified the need for improvements in managing unrealistic lab values and handling new drugs. Furthermore, using an LLM as an evaluator showed strong agreement with human assessments on relevance, completeness, and Prudence (>57%), demonstrating its potential to reduce human efforts. Benchmarking GPT-4o-based against LLaMA-based versions indicates the proposed evaluation framework's practical utility in making direct comparison across different ADS tools.

Conclusion. Our findings underscore the potential of the proposed evaluation framework in improving healthcare delivery of ADS tools while highlighting the need for robust governance to ensure safe, ethical integration. This work establishes a baseline for ADS tool evaluation and contributes to the discourse on AI governance in healthcare.

Improving Robustness of the Model-X Inference with Application to EHR Studies

♦*Molei Liu*

Peking University

The model-X conditional randomization test (CRT) is a flexible and powerful testing procedure for conditional independence testing. However, it requires perfect knowledge of the exposure X's conditional distribution and may lose its validity when there is an error in modeling X. This problem is even more severe when the adjustment covariates Z are high-dimensional. To address this challenge, we propose the Maxway CRT, which learns the conditional distribution of the response Y and uses it to calibrate the resampling distribution of X. We prove that the type-I error inflation of the Maxway CRT can be controlled by the learning error for a low-dimensional adjusting model plus the product of learning errors for X | Z and Y | Z, interpreted as an "almost doubly robust" property. Based on this, we develop implementing algorithms of the Maxway CRT in practical scenarios including surrogate-assisted semi-supervised learning and transfer learning. We apply our methodology to two real-world studies on electronic health record and biobank data.

25CHI055: New advances in design and analysis of longitudinal studies

Integrating multiple imperfect measures for alcohol use in longitudinal research studies

♦*Robert Cook, Samuel Wu, Donald Porchia, Yan Wang, Zhigang Li*

University of Florida, University of Florida, University of Florida, University of Florida, University of Florida

Alcohol consumption is common among adults with a chronic health condition. Researchers often use longitudinal data analyses to determine whether changes in alcohol consumption are associated with changes in health outcomes. Existing methods to measure alcohol consumption are each susceptible to measurement error. The goal of this study was to combine several individual measures of alcohol consumption into a single measure that would be more accurate than any of the individual measures.

Methods: We used data from 47 older adults who were participating in a clinical trial of an alcohol intervention. Alcohol was assessed by self-report (4 timepoints), a blood test for an alcohol metabolite called phosphatidylethanol (PEth), and a biosensor attached to the ankle (2 timepoints). Results from each alcohol measure were categorized into quintiles and also dichotomized as "heavy" or "not heavy" drinking during the previous 30 days. We calculated a latent score using the quintile data and determined a latent score cutpoint that maximized the accuracy of identifying heavy drinking.

Results: All participants had heavy drinking at baseline, and most reduced drinking over time. The latent score classification of heavy drinking was most strongly correlated with self-report (with zero false-positives compared to self-report), whereas the latent score provided 15% false-positives compared to PEth and 32% false-positives compared to the biosensor data. Overall accuracy comparisons were similar, with self-report having the highest maximum accuracy.

Conclusion. We successfully created a new"latent" variable that combined three imperfect measures of alcohol consumption. The next steps will be to assess whether the latent variable better correlates with expected clinical outcomes compared to the individual measures.

Integrating wearable sensors into longitudinal cohort studies: Opportunities and challenges

♦*Yan Wang*

University of Florida

Background: The ubiquity of smartphone and wearable mobile technologies provides a unique opportunity for researchers to collect real-time data embedded in classic longitudinal cohort studies. This talk will discuss the advantages and challenges with integrating wearable sensors into cohort study design using an ongoing cohort as an example.

Methods: The Study on Medical marijuana and Its Long-term Effects (SMILE study) is an ongoing longitudinal cohort study which recruits and follows older adults (50 years or older, 50% male) with chronic pain for one year. Some will have initiated medical marijuana (MM group) and others will not use MM (comparison group). Target enrollment will be 440 (3:1 ratio for MM and comparison group) participants. Collection of subjective and objective data occurs at in-person visits (baseline & 12 months) and via smartphone- and sensor-based measurement bursts at 1, 3, 6, 9, and 12 months. The smartphone-based ecological momentary assessment (EMA) data will capture detailed MM use patterns and subjective short-term real-time outcomes (e.g., pain intensity rating). These data will be complemented by objective real-time data collected using a wearable sensor-based Fitbit (e.g., physical activity and sleep).

Results: Data from wearable sensors provide valuable objective evidence on whether older adults who initiate MM have greater reductions in real-time improvements in physical functioning and sleep compared to those in the comparison group. However, integration and analysis of such data together with survey and EMA data could be challenging and may leads to inconsistent conclusions.

Conclusions: Integrating wearable sensors in classic cohort design may add valuable insights beyond self-reported surveys, but best practice needs to be established to fully utilize the rich data collected using these technology-based assessments.

Joint modeling in presence of informative censoring on the retrospective time scale

Quran Wu, Michael Daniels, Areej El-Jawahri, Marie Bakitas, ♦*Zhigang Li*

University of Florida, University of Florida, Harvard University, UAB, University of Florida

Joint modeling of longitudinal data such as quality of life data and survival data is important to draw efficient inferences because it can account for the associations between those two types of data. We develop a novel joint modeling approach that can address the challenge by allowing informative censoring events to be dependent on patients' quality of life and survival through a random effect. There are two sub-models in our approach: a linear mixed effect model for the longitudinal quality of life and a competing-risk model for the death time and dropout time that share the same random effect as the longitudinal model. Our approach can provide unbiased estimates for parameters of interest by appropriately modeling the informative censoring time. Model performance is assessed with a simulation study and compared with existing approaches. A real-world study is presented to illustrate the application of the new approach.

Correlation Coefficients for a Study with Repeated Measures

♦Guogen Shan

University of Florida

Repeated measures are increasingly collected in a study to investigate the trajectory of measures over time. One of the first research questions is to determine the correlation between two measures. The following five methods for correlation calculation are

compared: (1) Pearson correlation; (2) correlation of subject means; (3) partial correlation for subject effect; (4) partial correlation for visit eff5ect; and (5) a mixed model approach. Pearson correlation coefficient is traditionally used in a cross-sectional study.

Pearson correlation is close to the correlations computed from mixed-effects models that consider the correlation structure, but Pearson correlation may not be theoretically appropriate in a repeated-measure study as it ignores the correlation of the outcomes from multiple visits within the same subject. We compare these methods with regard to the average of correlation and the mean squared error. In general, correlation under the mixed-effects model with the compound symmetric structure is recommended as

its correlation is close to the nominal level with small mean square error.

25CHI071: Recent Advances in Nonparametric Estimation and Inference

Jackknife empirical likelihood for the correlation coefficient with multiplicative distortion measurement errors

Brian Pidgeon, Pangpang Liu, ♦Yichuan Zhao

Georgia State University, Purdue University, Georgia State University

In this paper, we consider the estimation problem of a correlation coefficient between two unobserved variables of interest that are distorted in a multiplicative way by some unobserved confounding variable. We investigate the direct plug-in estimator of the correlation coefficient. We propose using jackknife empirical likelihood (JEL) and its variations to construct confidence intervals for the correlation coefficient based on the estimator. The proposed JEL statistic is shown to be asymptotically a standard chi-squared distribution. We compare our methods to the previous empirical likelihood (EL) techniques of Zhang et al. (2014) and show the JEL possesses better small sample properties. Simulation studies are conducted to examine the performance of the proposed estimator, and we also use our proposed methods to analyse the Boston housing data for illustration.

Estimation of Multiple Large Precision Matrices and Its Application to High-Dimensional Quadratic Discriminant Analysis

Yilei Wu, Liyuan Zheng, ♦Yingli Qin, Mu Zhu, Weiming Li

University of Waterloo

When estimating precision matrices (the inverses of covariance matrices) for data from multiple related categories, such as subtypes of a particular disease, it is natural to expect that these precision matrices share some common structure. In this paper, we assume that the population precision matrix for each category can be decomposed into three components: a common diagonal component, a common low-rank component, and a category-specific low-rank component. This decomposition is motivated by latent factor models, where some latent factors have shared effects across all categories, while others exhibit category-specific effects.

We propose a method to jointly estimate these precision matrices, beginning with the estimation of the number of factors. Under mild conditions, we establish the consistency of our estimators. We then apply the proposed estimators to construct a high-dimensional quadratic discriminant analysis (QDA) classifier and derive its classification error convergence rate. Numerical examples are provided to illustrate the performance of our method.

A Unified Framework of Classification-based Equality Test of Distributions

Zhen Zhang, ♦Xin Liu

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

Testing the equality of two distributions based on two collected samples has been widely popular yet challenging. Traditional two-sample tests may deteriorate or fail in non-Gaussian settings or high dimensional spaces due to a lack of accurate estimation and relatively low test power. In this article, a unified model-free framework is proposed to test the equality of distributions using classifications, named the classification-based equality test of distributions (CETD). By training a fine-tuned classifier with two sample instances and sample labels, a test statistic is created as the estimated prediction error of the trained classifier. Asymptotic theories of the proposed test statistic are established under both null and alternative hypotheses, and the power of the test is investigated. Experiments on both synthetic and real datasets have demonstrated the excellent testing power of the proposed method.

Minimax optimal two-stage algorithm for moment estimation under covariate shift

Zhen Zhang, Xin Liu, ♦Shaoli Wang, Jiaye Teng

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

Covariate shift occurs when the distribution of input features differs between the training and testing phases. In covariate shift, estimating an unknown function's moment is a classical problem that remains under-explored, despite its common occurrence in real-world scenarios. In this talk, we investigate the minimax lower bound of the problem when the source and target distributions are known. To achieve the minimax optimal bound (up to a logarithmic factor), we propose a two-stage algorithm. Specifically, it first trains an optimal estimator for the function under the source distribution, and then uses a likelihood ratio reweighting procedure to calibrate the moment estimator. In practice, the source and target distributions are typically unknown, and estimating the likelihood ratio may be unstable. To solve this problem, we propose a truncated version of the estimator that ensures double robustness and provide the corresponding upper bound. Extensive numerical studies on synthetic examples confirm our theoretical findings and further illustrate the effectiveness of our proposed method.

25CHI072: Recent advances in privacy-protected data collection and analysis

Learning from vertically distributed data across multiple sites: An efficient privacy-preserving algorithm for Cox proportional hazards model with variable selection

Guanhong Miao, Lei Yu, Jingyun Yang, David Bennett, Jinying Zhao, ♦Samuel Wu

University of South Florida, Rush University Medical Center, Rush University Medical Center, Rush University Medical Center, University of South Florida, University of South Florida

Objective: To develop a lossless distributed algorithm for regularized Cox proportional hazards model with variable selection to support federated learning for vertically distributed data.

Methods: We propose a novel distributed algorithm for fitting regularized Cox proportional hazards model when data sharing among different data providers is restricted. Based on cyclical coordinate descent, the proposed algorithm computes intermediary statistics by each site and then exchanges them to update the model parameters in other sites without accessing individual patient-level data. We evaluate the performance of the proposed algorithm with (1) a simulation study and (2) a real-world data analysis predicting the risk of Alzheimer's dementia from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP). Moreover, we compared the performance of our method with existing privacy-preserving models.

Results: Our algorithm achieves privacy-preserving variable selection for time-to-event data in the vertically distributed setting, without degradation of accuracy compared with a centralized approach. Simulation demonstrates that our algorithm is highly efficient in analyzing high-dimensional datasets. Real-world data analysis reveals that our distributed Cox model yields higher accuracy in predicting the risk of Alzheimer's dementia than the conventional Cox model built by each data provider without data sharing. Moreover, our algorithm is computationally more efficient compared with existing privacy-preserving Cox models with or without regularization term.

Conclusion: The proposed algorithm is lossless, privacy-preserving and highly efficient to fit regularized Cox model for vertically distributed data. It provides a suitable and convenient approach for modeling time-to-event data in a distributed manner.

Distributed Proportional Likelihood Ratio Model With Application to Data Integration Across Clinical Sites

♦Jiasheng Shi, Chongliang Luo

The Chinese University of Hong Kong, Shenzhen, Washington University in St. Louis

Real-world evidence synthesis through the integration of data from distributed research networks has gained increasing attention in recent years. Due to privacy concerns and restrictions on sharing patient-level data, distributed algorithms that do not require sharing patient-level information are in great need of facilitating multisite collaborations. On the other hand, data collected at multiple sites often come from diverse populations, and there exists a substantial amount of heterogeneity across sites in patient characteristics. Most of the existing distributed algorithms have ignored such between-site heterogeneity. In this work, we aim to fill this methodological gap by proposing a general distributed algorithm. We develop our distributed algorithm based on a general semiparametric model, namely, the proportional likelihood ratio model, which is a semiparametric extension of the generalized linear model. We devise the proportional likelihood ratio model with site-specific baseline function, to account for between-site heterogeneity, and shared regression parameters to borrow information across sites. Under this flexible formulation, our distributed algorithm is designed to be privacy-preserving and communication-efficient (i.e., only one round of communication across sites is needed). We validate our method via simulation studies and demonstrate the utility of our method via a multisite study of pediatric avoidable hospitalization based on electronic health record data from a total of 354,672 patients across 26 different clinical sites within the Children's Hospital of Philadelphia health system.

Logistic Regression Model for Differentially-Private Matrix Masking Data

♦Linh Nghiem, Aidong Adam Ding, Samuel Wu

University of Sydney, Northeastern University, University of South Florida

A recently proposed scheme utilizing local noise addition and matrix masking enables data collection while protecting individual privacy from all parties, including the central data manager. Statistical analysis on such privacy-preserved data is particularly challenging for nonlinear models like logistic regression. By leveraging a relationship between logistic regression and linear regression estimators, we propose the first valid statistical analysis method for logistic regression under this setting. Theoretical analysis of the proposed estimators confirmed its validity under an asymptotic framework with increasing noise magnitude, which is uncommon in traditional measurement error model analysis, to account for strict privacy requirements. Simulations and real data analyses demonstrate the superiority of the proposed estimators over naive logistic regression methods on privacy-preserved data sets.

25CHI103: Statistical Learning for Complex Data Structures

Estimation and Inference for CP Tensor Factor Model

Bin Chen, ♦Yuefeng Han, Qiyang Yu

University of Rochester, University of Notre Dame, University of Rochester

High-dimensional tensor-valued data have recently gained attention from researchers in economics and statistics. We consider the estimation and inference of high-dimensional tensor factor models, where each dimension of the tensor diverges. Specifically, we focus on the factor model that admits CP-type tensor decomposition, allowing for loading vectors that are not necessarily orthogonal. Based on the contemporary covariance matrix, we propose an iterative higher-order projection estimation method. Our estimator is robust to weak dependence among factors and weak correlation across different dimensions in the idiosyncratic shocks. We develop an inferential theory, establishing consistency and the asymptotic normality under relaxed assumptions. Through a simulation study and an empirical application with sorted portfolios, we illustrate the advantages of our proposed estimator over existing methodologies in the literature.

Heteroskedastic Tensor Clustering

♦Yuchen Zhou, Yuxin Chen

University of Illinois Urbana-Champaign, University of Pennsylvania

Tensor clustering, which seeks to extract underlying cluster structures from noisy tensor observations, has gained increasing attention. One extensively studied model for tensor clustering is the tensor block model, which postulates the existence of clustering structures along each mode and has found broad applications in areas like multi-tissue gene expression analysis and multilayer network analysis. However, currently available computationally feasible methods for tensor clustering either are limited to handling i.i.d. sub-Gaussian noise or suffer from suboptimal statistical performance, which restrains their utility in applications that have to deal with heteroskedastic data and/or low signal-to-noise-ratio (SNR).

To overcome these challenges, we propose a two-stage method, named High-order HeteroClustering (HHC), which starts by performing tensor subspace estimation via a novel spectral algorithm called Thresholded Deflated-HeteroPCA, followed by approximate k-means to obtain cluster nodes. Encouragingly, our algorithm provably achieves exact clustering as long as the SNR exceeds the computational limit (ignoring logarithmic factors); here, the SNR refers to the ratio of the pairwise disparity between nodes to the noise level, and the computational limit indicates the lowest SNR that enables exact clustering with polynomial runtime. Comprehensive simulation and real-data experiments suggest that our algorithm outperforms existing algorithms across various settings, delivering more reliable clustering performance.

Interpretable Classification of Categorical Time Series Using the Spectral Envelope and Optimal Scalings

♦Zeda Li, Scott Bruce, Tian Cai

Baruch College, CUNY, Texas A&M University, The City University of New York

In this talk, I will introduce a novel approach to the classification of categorical time series under the supervised learning paradigm. To construct meaningful features for categorical time series classification, we consider two relevant quantities: the spectral envelope and its corresponding set of optimal scalings. These quantities characterize oscillatory patterns in a categorical time series as the largest possible power at each frequency, or spectral envelope, obtained by assigning numerical values, or scalings, to categories that optimally emphasize oscillations at each frequency. Our procedure combines these two quantities to produce an interpretable and parsimonious feature-based classifier that can be used to accurately determine group membership for categorical time series. Classification consistency of the proposed method is investigated, and simulation studies are used to demonstrate accuracy in classifying categorical time series with various underlying group structures. Finally, we use the proposed method to explore key differences in oscillatory patterns of sleep stage time series for patients with different sleep disorders and accurately classify patients accordingly.

Dynamic Supervised Principal Component Analysis for Classification

Wenbo Ouyang, ♦Ruiyang Wu, Ning Hao, Hao Zhang

University of Arizona, Baruch College, CUNY, University of Arizona, University of Arizona

In this talk, I will introduce a novel framework for high-dimensional dynamic classification to address the evolving nature of class distributions over time or other index variables. Under this framework, traditional discriminant analysis techniques are adapted to learn dynamic decision rules with respect to the index variable. It features a new supervised dimension reduction method employing kernel smoothing to identify the optimal subspace projection, and the resulting variables from this projection are trained with a subsequent classifier such as linear discriminant analysis and quadratic discriminant analysis. I will illustrate the effectiveness of the proposed methods through theoretical analysis and numerical examples. The results show considerable improvements in classification accuracy and computational efficiency. This work contributes to the field by offering a robust and adaptive solution to the challenges of scalability and non-staticity in high-dimensional data classification.

25CHI104: Statistical learning with complex data

Lessons learned from LLM benchmarking & evaluation

♦Youna Hu

Amazon

Large Language Model (LLM) evaluation has become increasingly critical as these models proliferate, yet establishing reliable benchmarking methodologies remains challenging. This talk presents a comprehensive analysis of LLM evaluation frameworks, examining both the fundamental questions we ask of these models and the metrics used to assess their responses. We explore five key model capability domains: instruction following, multi-turn interaction, long-context handling, agent behavior, and reasoning abilities, comparing performance across prominent closed-source models like ChatGPT and emerging open-source alternatives such as DeepSeek. Drawing from extensive benchmarking experience, we discuss crucial practical challenges including benchmark saturation, the limitations of public evaluation datasets, and the development of robust internal assessment frameworks. Special attention is given to the bidirectional relationship between evaluation metrics and model training, as well as the concerning lack of statistical rigor in current evaluation practices. This presentation aims to provide actionable insights for researchers and practitioners involved in LLM development and deployment, while highlighting critical areas for improvement in evaluation methodologies

Decentralized TD Learning with Spatio-temporal Information Dependence

♦Shaogao Lv

Nanjing Audit University

We are concerned with policy evaluation problems in distributed reinforcement learning (RL) by temporal-difference (TD) method with linear function approximation, including typical settings involved in multi-agent RL. In this talk, we develop a novel decentralized TD algorithm to estimate the value function of the multi-agent Markov decision process for a given target policy. At each agent, the direction of each update incorporates historical information about its own (pseudo)gradient, as well as information about the global gradient of the network at the current time. The method proposed in this talk for policy evaluation fully utilizes the space-time gradient information, which can improve the communication efficiency between agents. Theoretically, we provide a finite-time analysis for the proposed decentralized RL algorithm under both independent and identically distributed and Markovian data, respectively. Empirically, several simulation results with both linear approximation are presented to validate the proposed algorithm.

Quantile tensor factor regression with interaction effects and its application to multimodal data analysis

Pengfei Pi, ♦Shan Luo

Shanghai Jiao Tong University, Shanghai Jiao Tong University

Multimodal data analysis plays a pivotal role in advancing our comprehension of the brain, contributing significantly to fields such as neuroscience, psychology, psychiatry, and neurology. Multimodal data are consistently modeled as tensor covariates. For numerous clinical and psychological outcomes, quantile regression is highly valued due to its stability and flexibility. This article explores quantile regression involving tensor covariates and applies it to multimodal data analysis. In this proposed approach, tensor covariates are assumed to adhere to a factor structure, from which new feature variables are derived. Subsequently, the main effects of these new feature variables and their interactions are considered in the quantile regression. The article introduces a rapid and efficient method for predicting the dependent variable in quantile regression with interaction effects. The main theoretical results of our approach have been established. The accuracy and stability of the proposed algorithm are validated through extensive simulation experiments. Finally, the proposed method is applied to analyze SKCM and ADHD data, demonstrating superior predictive accuracy and faster computational speed compared to existing methods.

Robust group detection and membership prediction

Boyan Shen, ♦Xuerong Chen, Yong Zhou

Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, East China Normal University

In this paper, we propose a novel quantile regression modeling framework with a latent group structure, allowing samples drawn from a population consisting of groups with different conditional quantiles along with certain covariates. Different from most conventional modeling approaches for group identification, such as finite mixture models and threshold models, our new model is distribution free, allows the group numbers and group structure of regression coefficients to be the same or different for different covariates. We identify the potential group structure for the quantile regression coefficients using a regularization method and achieve group boundaries recovery through support vector machine (SVM) method, after artificially assigning appropriate labels to different groups. The computational burden of our approach is significantly lower than the pairwise fused regularization method. Moreover, unlike existing regularization methods, our method can analyze and explain the reasons for grouping and predict the group membership of new individuals based on the estimated group boundary. We establish the theoretical properties of the proposed estimators for group parameters and boundary parameters. Simulation studies and real data analysis illustrate that the proposed methods perform well.

25CHI039: Innovative Statistical and Machine Learning Methods for Complex Health Data

Transformer-Based Self-Supervised Learning for Multimodal Wearable Data

♦Jingjing Zou

UC San Diego

The use of wearable devices has become increasingly prevalent in recent years, enabling the collection of high-frequency data that provides valuable insights into individuals' health behavior patterns. With the rapid advancement of machine learning techniques, there has been a growing interest in leveraging these methods to extract meaningful latent representations from wearable device data and use the features extracted to classify types of activities and postures. This study explores the transformer-based autoencoders applied to the multiple-modality activity data collected by wearable devices. We proposes a novel approach to self-supervised learning of high-frequency wearable device data and evaluates its performance in capturing key features and patterns as well as in predicting posture types, compared to existing state-of-art approaches. The results demonstrate exceptional performance, supported by theoretical guarantees that provide a deeper understanding of the proposed approach.

Neyman Smooth-Type Goodness of Fit in Complex Surveys

Lang Zhou, ♦Yan Lu, Guoyi Zhang, Ronald Christensen

AbbVie Inc., University of New Mexico, University of New Mexico, University of New Mexico

In this research, we extend Neyman smooth-type goodness-of-fit tests to complex survey data by incorporating consistent estimators tailored to the survey design. This is achieved through data-driven nonparametric order selection methods. Simulation studies demonstrate that the proposed methods significantly enhance statistical power while effectively controlling the type I error rate, particularly when dealing with scenarios involving relatively constant probabilities. We further illustrate the application of these methods using data from the National Youth Tobacco Survey (NYTS).

Bayesian monotone regression with large number of covariates and complex structure

♦Ken Cheung, Keith Diaz

Columbia University, Columbia University

Making monotonicity assumption in regression, whenever reasonable, improves efficiency in estimation. However, when there are many covariates, performing monotone regression can be challenging as the number of computations grows exponentially with the size of the covariate space. We formulate the estimation of monotone response surface of multiple factors as the inverse of an iteration of partially ordered classifier ensemble. Each ensemble (called PIPE-classifier) is a projection of Bayes classifiers on the constrained space. We prove the inverse of PIPE-classifiers (iPIPE) exists and propose algorithms to efficiently compute iPIPE by reducing the space over which optimization is conducted. The methods are applied in analysis and simulation settings where the surface dimension is higher than what the monotone regression literature typically considers. Simulation shows iPIPE-based credible intervals achieve nominal coverage probability and are much more precise compared to unconstrained estimation.

Reference: Cheung and Diaz (2023). Monotone response surface of multi-factor condition: estimation and Bayes classifiers. JRSSB 85, 497-522.

Functional Fixed and Random Effects Inference with Applications to Accelerometry Data

♦*Erjia Cui*

University of Minnesota

We first introduce Fast Univariate Inference (FUI), a marginal approach to obtaining inferential results for functional fixed effects in functional mixed models. The approach consists of three steps: (1) fit massively univariate pointwise mixed-effects models; (2) apply any smoother along the functional domain; (3) obtain joint confidence bands using analytical or nonparametric approaches. Simulation studies show that model fitting and inference are accurate and much faster than existing approaches. To obtain uncertainty quantification for individual functions, we further extend FUI and introduce a functional random effects inference framework to obtain subject-specific inferential results, which combines local mixed models with global variance decomposition. Methods are applied to physical activity data measured by body-worn accelerometers collected from the National Health and Nutrition Examination Survey (NHANES).

25CHI053: Modern Statistical Modeling in Medical Research with Real World Data

Time-Since-Infection Model for Hospitalization and Incidence Data

Jiasheng Shi, Yizhao Zhou, ♦*Jing Huang*

The Chinese University of Hong Kong, Shenzhen, AstraZeneca, University of Pennsylvania

The Time since Infection (TSI) models have gained widespread popularity and acclaim for their performance during the COVID-19 pandemic, establishing them as an increasingly popular choice for modeling infectious disease transmission due to their practicality, flexibility and ability to address complex disease control questions. However, a notable limitation of TSI models is their primary reliance on incidence data. In existing TSI models, even when hospitalization data are available, they have not been designed to estimate disease transmission or to predict disease related hospitalizations — metrics crucial for understanding the trajectory of a pandemic and for hospital resource planning. Furthermore, their dependence on reported infection data makes them vulnerable to variations in data quality. In this study, we advance TSI models by integrating hospitalization data, a critical component for a comprehensive understanding of infectious diseases. Our improvement enables hospitalization modeling, reduces bias in incidence data, and connects TSI models with other infectious disease models. We introduce hospitalization propensity parameters to model incidence and hospitalization counts jointly. We use a composite likelihood function to accommodate complex data structures and an MCEM algorithm to effectively estimate model parameters. We apply our method to COVID-19 data and estimate disease transmission dynamics, assess risk factor impacts, and calculate hospitalization propensities. Our novel TSI model offers a fresh perspective on using hospitalization data to enhance the understanding of disease dynamics and support public health efforts.

Causal inference for all: Estimands of practical interest in intercurrent event settings

♦*Ruixuan Zhao, Linbo Wang, Mats J. Stensrud*

University of Toronto Scarborough, University of Toronto Scarborough, Ecole Polytechnique Fédérale de Lausanne

Researchers often express interest in treatment effects that adequately account for post-treatment events (so-called intercurrent events). However, outcome contrasts that naively condition on intercurrent events lack a straightforward causal interpretation, and the practical relevance of other commonly used approaches is debated. In this work, we propose strategies for formulating and choosing an estimand, beyond the marginal intention-to-treat effect, from the perspective of decision-makers and treatment developers. We emphasize that a well-articulated, practically useful research question should either reflect decision-making at this point in time or future treatment development. A common feature of estimands that are practically useful is their correspondence to possibly hypothetical but well-defined interventions in identifiable (sub)populations. To illustrate our points, we consider examples that have recently motivated the consideration of principal stratum estimands in clinical trials. In all of these examples, we suggest alternative causal estimands that align with explicit research questions of practical interest and require less stringent identification assumptions.

A functional spatial partitioning approach to lesion segmentation using MRIs

♦*Lin Zhang*

University of Minnesota

Imaging plays an important role in cancer diagnosis and staging by noninvasively evaluating the presence and extent of local and distant disease. Computer aided detection algorithms are being developed for fast and reproducible cancer diagnosis from complex and high-dimensional medical imaging data. While extensive statistical methods have been developed for voxel-wise cancer classification, existing lesion segmentation methods primarily rely on deep learning methods centered at the convolutional neural networks. We have developed a functional spatial partitioning approach for automatic lesion-wise cancer detection using imaging data, which jointly estimate the lesion boundaries and the spatial processes within each partitioned region in a Bayesian framework.

25CHI054: New Advancements in Statistical Learning

Neural network on interval-censored data with application to the prediction of Alzheimer's disease

♦*Tao Sun, Ying Ding*

Renmin University of China, University of Pittsburgh

Alzheimer's disease (AD) is a progressive and polygenic disorder that affects millions of individuals each year. Given that there have been few effective treatments yet for AD, it is highly desirable to develop an accurate model to predict the full disease progression profile based on an individual's genetic characteristics for early prevention and clinical management. This work uses data composed of all four phases of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, including 1740 individuals with 8 million genetic variants. We tackle several challenges in this data, characterized by large-scale genetic data, interval-censored outcome due to intermittent assessments, and left truncation in one study phase (ADNIGO). Specifically, we first develop a semiparametric transformation model on interval-censored and left-truncated data and estimate parameters through a sieve approach. Then we propose a computationally efficient generalized score test to identify variants associated with AD progression. Next, we implement a novel neural network on interval-censored data (NN-IC) to construct a prediction model using top variants identified from the genome-wide test. Comprehensive simulation studies show that the NN-IC outperforms several existing methods in terms of prediction accuracy. Finally, we apply the NN-IC to the full ADNI data and successfully identify subgroups with differential progression risk profiles.

Variational Bayesian Semi-supervised Keyword Extraction

♦*Yaofang Hu, Yichen Cheng, Yusen Xia, Xinlei Wang*

University of Alabama, Georgia State University, Georgia State University, University of Texas at Arlington

The expansion of textual data, stemming from various sources such as online product reviews and scholarly publications on scientific discoveries, has created a demand for the extraction of succinct yet comprehensive information. As a result, in recent years, efforts have been spent in developing novel methodologies for keyword extraction. Although many methods have been proposed to automatically extract keywords in the contexts of both unsupervised and fully supervised learning, how to effectively use a partial list of keywords, such as author-specified keywords and Twitter hashtags, remains an under-explored area. We propose a novel variational Bayesian semi-supervised (VBSS) keyword extraction approach, built on a recent Bayesian semi-supervised (BSS) technique that uses the information from a small set of known keywords to identify previously undetected ones. Our proposed VBSS method greatly enhances the computational efficiency of BSS via mean-field variational inference, coupled with data augmentation, which brings closed-form solutions at each step of the optimization process. Further, our numerical results show that VBSS method offers enhanced performance for long texts and improved control over false discovery rates when compared with a list of state-of-the-art keyword extraction methods.

Diffusion model for large spatial temporal data

♦*Xin Tong*

National University of Singapore

Diffusion model (DM) is a novel computational method for sample generation. It is useful for Bayesian statistics due to its fast computational speed and the potential for prior distribution modeling. However, for high dimensional distributions, the training cost of DM is subject to the curse of dimensionality. In this talk, we will leverage the sparse local structure that can commonly be found in spatial temporal data to design DM training strategy that can avoid the curse of dimension. The main analysis technique relies on a localized Stein method.

Consistent Order Determination of Markov Decision Process

♦*Chuyun Ye, Lixing Zhu, Ruoqing Zhu*

Beijing Normal University, Beijing Normal University at Zhuhai, University of Illinois at Urbana-Champaign

Reinforcement learning (RL) leverages the Markov Decision Process (MDP), which fundamentally relies on the Markov property. However, numerous real-world systems exhibit extended temporal dependencies, demanding higher-order Markov models beyond the typical first-order assumption. This paper tackles the challenge of consistently estimating the order of such Markov processes, a problem where traditional sequential testing methods are hindered by limitations in sensitivity and consistency. We introduce a novel, two-stage estimation procedure: first, we define a function that precisely captures the k-order Markov assumption, guaranteeing sensitivity to all violations; second, we construct a signal statistic that consistently identifies the true order by exploiting a distinct pattern of minimizers. This approach yields a consistent estimator and facilitates efficient implementation. Furthermore, the characteristic curve pattern of the signal statistic aids in visual inspection, that could simplify the order determination process in practical applications. We validate the effectiveness of our method through simulations and a real-world dataset, representing a significant stride in accurately modeling and applying RL to systems with complex temporal dependencies.

25CHI058: New statistical methods in nonlinear regression analyses

Regression Analysis of Semiparametric Cox-Aalen Transformation Models with Partly Interval-Censored Data

Xi Ning, ♦*Yanqing Sun, Yinghao Pan, Peter Gilbert*

Colby College, University of North Carolina at Charlotte, University of North Carolina at Charlotte, University of Washington and Fred Hutchinson Cancer Center

Partly interval-censored data, comprising exact and interval-censored observations, are prevalent in biomedical, clinical, and epidemiological studies. This paper studies a flexible class of the semiparametric Cox-Aalen transformation models for regression analysis of such data. These models offer a versatile framework by accommodating both multiplicative and additive covariate effects and both constant and time-varying effects within a transformation, while also allowing for potentially time-dependent covariates. Moreover, this class of models includes many popular models such as the semiparametric transformation model, the Cox-Aalen model, the stratified Cox model, and the stratified proportional odds model as special cases. To facilitate efficient computation, we formulate a set of estimating equations and propose an Expectation-Solving (ES) algorithm that guarantees stability and rapid convergence. Under mild regularity assumptions, the resulting estimator is shown to be consistent and asymptotically normal. The validity of the weighted bootstrap is also established. A supremum test is proposed to test the time-varying covariate effects. Finally, the proposed method is evaluated through comprehensive simulations and applied to analyze data from a randomized HIV/AIDS trial. This is a joint work with Xi Ning, Yinghao Pan and Peter B. Gilbert.

Collaborative quantile treatment effect estimation

Ye Fan, Ying Wei, Sung Nok Chiu, Tiejun Tong, ♦Nan Lin

Capital University of Economics and Business, Columbia University, Hong Kong Baptist University, Hong Kong Baptist University, Washington University in St. Louis

Heterogeneous treatment effect estimation plays a pivotal role in personalized medicine. Quantile treatment effect (QTE) estimation provides a more nuanced perspective by capturing heterogeneity across subgroups at different quantile levels of the outcome distribution. Despite well-established statistical theory, the computation of QTE estimation in distributed big data presents significant challenges. In this work, we introduce SCQTE, a novel sequential collaborative method designed to facilitate efficient QTE estimation in distributed data environments. SCQTE offers a unified and computationally efficient framework that accommodates various types of QTE estimation, including conditional and unconditional QTE. Under mild conditions, we establish that the SCQTE estimator is consistent and asymptotically equivalent to the oracle estimator derived from the full dataset.

clusterMLD: An Efficient Clustering Method for Multivariate Longitudinal Data

Junyi Zhou, ♦Ying Zhang, Wenzhou Zhou

Amgen Inc., University of Nebraska Medical Center, Indiana University

Longitudinal data clustering is challenging, especially with sparse and irregular observations. The literature lacks reliable methods dealing with clustering complicated longitudinal data, particularly with multiple longitudinal outcomes. In this manuscript, a new agglomerative hierarchical clustering method is developed in conjunction with B-spline curve fitting and constructing a unique dissimilarity measure for differentiating longitudinal observations. In an extensive simulation study, the proposed method demonstrates its superior performance in clustering accuracy and numerical efficiency compared to the existing methods. Moreover, the method can be easily extended to multiple-outcome longitudinal data without too much cost in computation and shows robust results against the complexity of the underlying mixture of longitudinal data. Finally, the method is applied to a data set from the SPRINT Study for validating the intervention efficacy in a Systolic Blood Pressure Intervention Trial and to a 12-year multi-site observational study (PREDICT-HD) for identifying the disease progression patterns of Huntington's disease (HD).

Joint mixed membership modeling of multivariate longitudinal and survival data

Yuyang He, ♦Xinyuan Song, Kai Kang

The Chinese University of Hong Kong, The Chinese University of Hong Kong, Sun Yat-sen University

This study presents a novel joint mixed membership model for multivariate longitudinal AD-related biomarkers and time of AD diagnosis. Unlike conventional finite mixture models that assign each subject a single subgroup membership, the proposed model assigns partial membership across subgroups, allowing subjects to lie between two or more subgroups. This flexible structure enables individualized disease progression and facilitates identifying clinically meaningful neurological statuses often elusive in current mixed effects models. We employ a spline-based trajectory model to characterize complex and possibly nonlinear patterns of multiple longitudinal clinical markers. A Cox model is then used to examine the effects of time-variant risk factors on the hazard of developing AD. We develop a Bayesian method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical inference. The proposed approach is assessed through extensive simulation studies and an application to the Alzheimer's Disease Neuroimaging Initiative study, showing a better performance in AD diagnosis than existing joint models.

25CHI060: Novel statistical methods for complex data analysis

A Joint Model for Multiple Longitudinal Data with Different Missing Data Patterns and with Applications to HIV Prevention Trials

♦Jing Wu, Ming-Hui Chen, Jeffrey Fisher

University of Rhode Island, University of Connecticut, University of Connecticut

In longitudinal clinical trials, it is common that mixed types of outcomes are collected on the same subject over time. It is also routinely encountered that all outcomes may be subject to substantial missing values due to dropout and intermittent missingness. Additionally, the missing data patterns of the mixed types of outcomes are usually the same for dropout while different for intermittent missingness. In this paper, a sequential multinomial model is adopted for dropout and subsequently, a new joint conditional model is constructed for intermittent missingness of mixed types of outcomes. The new model captures the complex structure of missingness and incorporates dropout and different intermittent missingness simultaneously. A bivariate zero-inflated Poisson mixed-effects regression model is assumed for the longitudinal count response data, respectively. We further show that the joint posterior distribution is improper if uniform priors are specified for the regression coefficients under the proposed model. An efficient Gibbs sampling algorithm is developed using a hierarchical centering technique. A modified logarithm of the pseudomarginal likelihood (LPML) and a new concordance measure criterion are used to compare the models under different missing data mechanisms. An extensive simulation study is conducted to investigate the empirical performance of the proposed methods, and the methods are further illustrated using real data from an HIV prevention clinical trial.

Model identification and selection for varying-coefficient EV models with missing responses

♦Mingtao Zhao, Houwu Wu, Fanqun Li

Anhui University of Finance and Economics, Anhui University of Finance and Economics, Anhui University of Finance and Economics

In this paper, I will introduce a model identification and selection approach for varying-coefficient EV models with missing responses, termed the bias-corrected double-penalized estimating equations (bcDPEE) method. The proposed approach does not need to assume whether the regression coefficients are constants or varying coefficients. First, it utilizes B-spline basis functions to approximate the nonparametric regression coefficients. Subsequently, the bcDPEE is constructed based on the observed responses , while accounting for the bias in the unobserved covariates. The missing responses are then imputed via the kernel estimation technique. Lastly, the bcDPEE is constructed to do model identification, estimation and selection simultaneously. Under some regularity conditions, the proposed approach consistently identifies and selects nonzero constant coefficients and varying coefficients. Moreover, the estimators of the varying coefficients achieve the optimal convergence rate for nonparametric function estimation, and the estimators of the nonzero constant coefficients have consistency and asymptotic normality. The finite-sample performance of the proposed method is evaluated through mento simulations and real data analysis.

A Self-Normalized Two-Sample Test for Nonaligned Time Series with Heavy Tails and Long Memory

Weiliang Wang, Yu Shao, ♦ *Ting Zhang*

Boston University, Boston University, University of Georgia

We consider the problem of two-sample testing for nonaligned time series with heavy tails and long memory, for which there are two major challenges. The first is to deal with the nonstandard convergence rate and limiting distribution, which often depend on the unknown degree of heavy-tailedness and long-range dependence. In addition, the two time series to be compared may exhibit different degrees of heavy-tailedness and long-range dependence, and in practice it can be difficult to determine if they share the same convergence rate or if one dominates the other. The second challenge is to handle nonaligned time series data, where the two time series to be compared may be observed over different time periods with different lengths. This requires a careful consideration on the joint dependence of overlapping and nonoverlapping segments when designing a valid statistical inference procedure. We illustrate how the technique of self-normalization can be adapted with a time warp to address the aforementioned challenges and lead us to a convenient and rigorous statistical method for two-sample testing of nonaligned time series with heavy tails and long memory. Numerical experiments including a Monte Carlo simulation study and real data applications are also provided.

Neural frailty machines for survival analysis

Jiawei Qiao, Ruofan Wu, Guanhua Fang, ♦*Wen Yu, Zhiliang Ying*

Fudan University, Ant Group, Fudan University, Fudan University, Columbia University

We propose a powerful and flexible neural modeling framework for survival regression. The framework basically assumes a separated structure of the baseline hazard rate and the nonlinear covariates effect. Meanwhile, a multiplicative frailty is introduced to capture the unobserved heterogeneity among individuals and the deep neural network architectures are adopted to approximate the baseline hazard rate and the nonlinear covariate structures, leading the proposed framework called neural frailty machines (NFM). The NFM can be viewed as an extension of neural proportional hazard models and includes many commonly used survival regression models as special cases. The likelihood function for right censored data is used to serve as the objective. The proposed algorithm allows efficient stochastic training, which can easily scale to large datasets. The estimation accuracy is measured by a metric defined through a Hellinger-type distance for hazard rate function. The non-asymptotic bounds for the estimation errors based on the Hellinger-type distance are derived. Then the consistency of the proposed neural estimators is established and the convergence rates are obtained. The rates are shown to reach the optimal speed of nonparametric regression estimation. Some simulation studies are carried out to verify the theoretical findings. The prediction performance of the proposed NFM models is evaluated over 6 benchmark datasets with different scales. The results show evidence on the improvement of the proposed method compared with the existing state-of-the-art survival models.

25CHI077: Recent Advances in Statistical Learning for Biological and Biomedical Data

Contrastive Learning on Multimodal Analysis of Electronic Health Records

♦*Doudou Zhou, Tianxi Cai, Feiqing Huang, Ryumei Nakada, Linjun Zhang*

National University of Singapore, Harvard T.H. Chan School of Public Health, Harvard T.H. Chan School of Public Health, Rutgers University, Rutgers University

Electronic health record (EHR) systems contain a wealth of multimodal clinical data including structured data like clinical codes and unstructured data such as clinical notes. However, many existing EHR-focused studies has traditionally either concentrated on an individual modality or merged different modalities in a rather rudimentary fashion. This approach often results in the perception of structured and unstructured data as separate entities, neglecting the inherent synergy between them. Specifically, the two important modalities contain clinically relevant, inextricably linked and complementary health information. A more complete picture of a patient's medical history is captured by the joint analysis of the two modalities of data. Despite the great success of multimodal contrastive learning on vision-language, its potential remains under-explored in the realm of multimodal EHR, particularly in terms of its theoretical understanding. To accommodate the statistical analysis of multimodal EHR data, in this paper, we propose a novel multimodal feature embedding generative model and design a multimodal contrastive loss to obtain the multimodal EHR feature representation. Our theoretical analysis demonstrates the effectiveness of multimodal learning compared to single-modality learning and connects the solution of the loss function to the singular value decomposition of a pointwise mutual information matrix. This connection paves the way for a privacy-preserving algorithm tailored for multimodal EHR feature representation learning. Simulation studies show that the proposed algorithm performs well under a variety of configurations. We further validate the clinical utility of the proposed algorithm in real-world EHR data.

Convex Covariate-adjusted Gaussian Graphical Regression

Ruobin Liu, ⬧Guo Yu

University of California, Santa Barbara

University of California, Santa Barbara

Gaussian graphical models (GGMs) are widely used for recovering the conditional independence structure among random variables. Recently, several key advances have been made to exploit an additional set of variables for better estimating the GGMs of the variables of interest. For example, in co-expression quantitative trait locus (eQTL) studies, both the mean expression level of genes as well as their pairwise conditional independence structure may be adjusted by genetic variants local to those genes. Existing methods to estimate covariate-adjusted GGMs either allow only the mean to depend on covariates or suffer from poor scaling assumptions due to the inherent non-convexity of simultaneously estimating the mean and precision matrix. We propose a convex formulation that jointly estimates the covariate-adjusted mean and precision matrix by utilizing the natural parametrization of the multivariate Gaussian likelihood. We verify our theoretical results with numerical simulations and perform a reanalysis of an eQTL study of glioblastoma multiforme (GBM), an aggressive form of brain cancer.

Instrumental variable analysis with multivariate point process treatments.

Yu Liu, Zhichao Jiang, ⬧Shizhe Chen

University of North Carolina, Chapel Hill, Sun Yat-Sen University, University of California, Davis

Multivariate point processes are popular tools for inferring relationships among subjects from recurrent event data such as neural spike trains. Complicated by the unmeasured confounding variables, interventions to the system are often employed in order to infer causality. However, these interventions are of low precision that they might influence the intensities of multiple processes simultaneously. In this study, we propose an instrumental variable framework with treatments being multivariate point processes. We show that the causal effects can be learned using generalized Wald estimation. We propose a penalized estimation procedure motivated by classic methods for density deconvolution. The proposed method is applied to neural data from behavioral experiments on mice.

25CHI107: Statistical methods for the analysis of complex data

Targeted Inference for High-Dimensional Quantile Regression Models

Yakun Liang, Xuejun Jiang, ⬧Jiancheng Jiang

Southern University of Science and Technology, Southern University of Science and Technology, University of North Carolina at Charlotte

This research introduces an innovative inference framework that employs dimension reduced convolution-smoothed quantile regression, while avoiding estimating the inverse of high-dimensional covariance matrix of the predictors. By calibrating the regularization parameter, we develop a data-driven test that can be shown to be an oracle test with probability tending to one. To mitigate the selective bias induced by dimension reduction and ensure valid inference, we implement a cross-fitting strategy by dividing the dataset into two parts: one for model selection and the other for parameter estimation. This process yields a fused estimator, derived from an informative weighting method that combines estimators from both dataset partitions. The fused estimator aids in constructing confidence intervals and performing Wald-type tests for targeted parameters. We establish the Bahadur representation of this estimator and obtain limiting distributions of the test statistics under both null and alternative hypotheses, with the number of parameters diverging to infinity. Advantages of our tests are further highlighted by theoretical power comparisons to some competitive tests. Empirical studies confirm effectiveness of the proposed tests across various linear parameter hypotheses. Additionally, we illustrate the use of the proposed methodology through two real-world data analyses.

Semiparametric Accelerated Failure Time Cure Model for Clustered Survival Data

Yi Niu, Duze Fan, Jie Ding, ⬧Yingwei Peng

Dalian University of Technology, Dalian University of Technology, Dalian University of Technology, Queen's University

The semiparametric accelerated failure time mixture cure model is an appealing alternative to the proportional hazards mixture cure model in analyzing failure time data with long-term survivors. However, this model was only proposed for independent survival data and it has not been extended to clustered or correlated survival data, partly due to the complexity of the estimation method for the model. In this talk, we present a marginal semiparametric accelerated failure time mixture cure model for clustered right-censored failure time data with a potential cure fraction. We overcome the complexity of the existing semiparametric method by proposing a generalized estimating equations approach based on the EM algorithm to estimate the regression parameters in the model. The correlation structures within clusters are modeled by working correlation matrices in the proposed generalized estimating equations. The large sample properties of the regression estimators are established. Numerical studies demonstrate that the proposed estimation method is easy to use and robust to the misspecification of working matrices and that higher efficiency is achieved when the working correlation structure is closer to the true correlation structure. We apply the proposed model and estimation method to a contralateral breast cancer study and reveal new insights when the potential correlation between patients is taken into account.

Robust causal effect estimation in high dimensional survival analysis via nonparametric learning

⬧Shanshan Ding, Zhezhen Jin

University of Delaware, Columbia University

Causal inference is an important tool to make inferences about causal relationships between variables. In this talk, we develop a robust nonparametric framework for estimating causal treatment effect in high dimensional survival analysis. The proposed framework is very flexible and alleviates assumptions in existing causal survival analysis. It allows the covariate dimension to grow exponentially fast with the sample size, without imposing stringent model or distributional assumptions for estimating the causal effect. The effectiveness of the methods will be demonstrated through both theoretical and numerical studies.

On Data-Enriched Logistic Regression

Cheng Zheng, Sayan Dasgupta, Yuxiang Xie, Asad Haris, ♦Yingqing Chen

University of Nebraska Medical Center, Fred Hutchinson Cancer Center, University of Washington, University of Washington, Stanford University

Biomedical researchers typically investigate the effects of specific exposures on disease risks within a well-defined population. The gold standard for such studies is to design a trial with an appropriately sampled cohort. However, due to the high cost of such trials, the collected sample sizes are often limited, making it difficult to accurately estimate the effects of certain exposures. In this paper, we discuss how to leverage the information from external "big data" (datasets with significantly larger sample sizes) to improve the estimation accuracy at the risk of introducing a small amount of bias. We propose a family of weighted estimators to balance bias increase and variance reduction when incorporating the big data. We establish a connection between our proposed estimator and the well-known penalized regression estimators. We derive optimal weights using both second-order and higher-order asymptotic expansions. Through extensive simulation studies, we demonstrate that the improvement in mean square error (MSE) for the regression coefficient can be substantial even with finite sample sizes, and our weighted method outperformed existing approaches such as penalized regression and James–Stein estimator. Additionally, we provide a theoretical guarantee that the proposed estimators will never yield an asymptotic MSE larger than the maximum likelihood estimator using small data only in general.

Abstracts of Contributed Sessions

Group 1: Bayesian Methods and Applications

Bayesian analysis of Cox-type regression model with partly linear covariate effects via reversible jump Markov chain Monte Carlo

♦Hengtao Zhang, Yuanke Qu, Kin Yau Wong, Chun Yin Lee

Guangdong Ocean University, Guangdong Ocean University, Hong Kong Polytechnic University, Hong Kong Polytechnic University

The partly linear Cox-type regression model provides a robust and flexible frame-work for exploring the potential nonlinear effects of a continuous variable on a censored outcome within the context of complex diseases. Most contemporary research studies the estimation methods of such a model based on the Frequentist paradigm, necessitating the selection of bandwidth and/or the number of spline basis functions for smoothing. We propose a Bayesian estimation approach that eliminates this requirement by employing the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. The proposed method can inherently estimate both the number and location of knots in the unknown function in a data-adaptive manner during the posterior inference process, rendering its applicability and predictive accuracy. We evaluate the finite-sample performance of the proposed method through a simulation study. The effectiveness of the proposed method is illustrated through the analysis of two medical datasets.

Bayesian crossover trial with binary data and extension to Latin-square design

♦Mingan Yang

University of New Mexico

In clinical trials, crossover design is widely used to assess treatment effects of drugs. Due to many practical issues, each patient in the study may receive only a subset of treatments under comparison, which is called an incomplete block crossover design. Correspondingly,the associated challenges are limited information and small sample size. In addition, the outcome is binary instead of continuous. In this article, we propose a Bayesian approach to analyze the crossover design with binary data. Markov chain sampling method is used to analyze the model and explore the extension to Latin-square design. We use several approaches such as data augmentation, scaled mixture of normal representation, parameter expansion to improve efficiency. The approach is illustrated using a simulation study and a real data example.

Birth-Death MCMC for Bayesian Variable Selection in Polygenic Risk Score Models

♦Nanwei Wang

University of New Brunswick

The Birth-Death Markov Chain Monte Carlo (BDMCMC) algorithm is a stochastic process that operates in continuous time, designed for Bayesian variable selection in polygenic risk score models. It dynamically navigates the model space by adding (birth) and removing (death) SNPs based on their posterior model probabilities. Unlike traditional discrete-time MCMC methods, BDMCMC assigns probabilistic weights to the sampled models through waiting times, making it a more efficient and scalable option for posterior inference. The algorithm's theoretical foundations are discussed, including its stationary distribution and birth-death rate functions. A detailed step-by-step procedure is provided, covering everything from model initialization to estimating posterior inclusion probabilities, showcasing how BDMCMC serves as a systematic and computationally feasible method for selecting SNPs in complex genomic datasets.

Tree-Based Bayesian Methods for Analyzing Partially Ordered Latent Statuses of Parkinson's Disease

♦Zhihao Wu,Xinyuan Song,Kai Kang,

The Chinese University of Hong Kong, The Chinese University of Hong Kong, Sun Yat-sen University

Medical data with partially ordered structures has been extensively utilized in diagnosing and predicting Parkinson's disease (PD). However, there has been a notable oversight in addressing the complex relationships between multiple biomarkers and complex latent disease progression patterns. This paper introduces a novel Bayesian framework that integrates Cognitive Diagnostic Models (CDMs) and Bayesian Additive Regression Trees (BART) to elucidate these intricate relationships, thereby facilitating a better understanding of individualized PD progression. The framework commences with mapping binary attribute vectors to partially ordered classes, extending traditional CDM approaches. Subsequently, BART is implemented without specific likelihood assumptions, accommodating complex relationships among variables while handling missing data through a unified Bayesian approach. An novel Metropolis-Hastings algorithm is developed for updating tree structures and parameters in the Bayesian inference process. Extensive simulation studies demonstrate the efficacy of our methodology. The framework is applied to the Parkinson's Progressive Markers Initiative (PPMI) dataset, yielding insights into disease partially ordered status and identifying significant biomarkers. This methodology offers a promising approach for analyzing complex medical data with latent structures, providing valuable insights for clinical applications in PD research.

Group 2:High-Dimensional Inference

A General U-Statistic Framework for High-Dimensional Multiple Change-Point Analysis

♦*Bin Liu,Yufeng Liu*

Fudan University, University of North Carolina at Chapel Hill, U.S.A

High-dimensional change-point analysis is essential in modern statistical inference. However, existing methods are often designed either for specific parameters (e.g., mean or variance) or for particular tasks (e.g., testing or estimation), making them difficult to generalize. Moreover, they typically rely on restrictive distributional assumptions, limiting their robustness to heavy-tailed data. We propose a unified framework for testing, estimating, and inferring multiple change points in high-dimensional data. Our approach leverages a two-sample U-statistic within a moving window, allowing flexible kernel function selection to accommodate structural changes in general parameters. For testing, we develop an $L\infty$-norm-based statistic with a high-dimensional multiplier bootstrap, achieving minimax-optimal power under sparse alternatives. For estimation, we construct an initial estimator for change-point number and locations and refine it using the U-statistic Projection Refinement Algorithm(U-PRA), attaining minimax-optimal localization rates. We further derive the asymptotic distribution of refined estimators, enabling valid confidence interval construction. Extensive numerical experiments demonstrate the superior performance of our method across various settings, including heavy-tailed distributions. Applications to genomic copy number variation and financial time series data highlight its practical utility.

Quadratic form estimation for moderate-dimensional logistic regression models

♦*Lingfeng Lyu, Xiao Guo*

University of Science and Technology of China, University of Science and Technology of China

Statistical inferences for quadratic form of logistic regression parameter have found wide applications. Classical theory based on Maximum likelihood estimator works perfectly in the low-dimensional regime, but fails when the parameter dimension p_n grows proportionally to the sample size n. We focus on moderate-dimensional logistic regression where n and p_n become increasingly large in a fixed ratio without imposing sparsity assumption on the regression parameter. We propose a novel estimator for the quadratic forms of the regression parameter based on the MLE and the bias and variance corrected procedure. The performance of the proposed method is illustrated both theoretically and numerically.

Randomized empirical likelihood test]{Randomized empirical likelihood test for ultra-high dimensional means under general covariances

♦*Yuexin Chen,Lixing Zhu,Wangli Xu,*

Renmin University of China, Beijing Normal University, Renmin University of China

This paper proposes a calibrated empirical likelihood test for ultra-high dimensional means that incorporates multiple projections. Under weak moment conditions on the distributions of data, we analyse all possible asymptotic distributions of the proposed test statistic in different scenarios. To determine the critical value and enhance test power, we employ the random symmetrization method based on the group of sign flips and use multiple selected projections. The test can still maintain the significance level asymptotically, even in the presence of heterogeneity in the data distribution. Moreover, the proposed test procedure allows for general covariance structures and ultra-high dimensional regimes. Further, the power function reveals the relation with the projection term in an asymptotic sense such that we can select suitable projections to achieve good power in various scenarios. A quasi-Newton algorithm is introduced to reduce the computational cost arising from the intensive optimizations required for computing empirical likelihood. Numerical studies evidence the promising performance of the proposed test compared with existing tests.

Debiased Inference for High-Dimensional Censored Quantile Regression via L1 Penalization

♦*Yu Guo,Tony Sit*

The Chinese University of Hong Kong, The Chinese University of Hong Kong

This paper introduces a novel methodology for constructing confidence intervals in high-dimensional censored quantile regression, where the number of covariates may substantially exceed the sample size. Building upon the weighted loss function proposed by Wang and Wang (2009), we incorporate an L1 penalty to handle high dimensionality and apply a debiasing procedure to correct the inherent bias introduced by the LASSO estimator. The resulting debiased estimator is shown to be asymptotically normal, forming a solid foundation for valid statistical inference. Notably, our approach relaxes the conventional global linearity assumption to a local linearity condition near the quantile of interest, enhancing model flexibility and robustness—especially in the presence of heteroskedasticity or violations of global linear effects. Simulation studies demonstrate the superior performance of our method in terms of coverage accuracy and efficiency when constructing confidence intervals, compared to existing approaches.

Group 3: Statistical Tests & Diagnostics

Significance test and application of principal components

♦*Lin Haiming,Shi Li*

The superiority criterion and significance test of principal component analysis have attracted the attention and research of professors Bartlett, Anderson, Johnson and Wichern. According to the test elements and their framework, this paper puts forward seven problems to be solved, and discusses them one by one. Firstly, based on the principal component model and its solution, an extended model including common components, special components and errors was constructed. Second, the excellent criterion of principal component analysis was established. Thirdly, the following five tests of principal components are proposed: simple structure test of correlation matrix between variables and principal components, special component test, common component test, common variance test, principal component naming test and correction measures, which solve the seven problems proposed here. Finally, the application steps of principal component significance test are elaborated with examples.

Influence Diagnostics for Generalized CP Tensor Regression Models

Chengcheng Hao, ♦Shaoyun Zhang, Yonghui Liu, Shuangzhe Liu

Shanghai University of International Business and Economics, Shanghai University of International Business and Economics, Shanghai University of International Business and Economics, University of Canberra

Tensor regression models have increasingly found applications across diverse fields. However, influence diagnostics within these models remain underexplored. This study extends local influence analysis and the case-deletion method to the generalized CP tensor model. One-step approximations of generalized Cook's distance are derived using Hessian and Fisher information matrices. Three perturbation schemes (i.e. case-weighted, single-explanatory-variable, and group-explanatory-variable) are analyzed via the largest curvature of likelihood displacement. Simulations and empirical analyses demonstrate that group-explanatory-variable perturbation diagnostics slightly outperform the other schemes.

Smooth Tests for Normality in ANOVA

♦Peiwen Jia,Xiaojun Song,Haoyu Wei,

Peking University, Peking University, University of California

The normality assumption for errors is fundamental in the analysis of variance (ANOVA) models, yet it is rarely formally tested in practice. In this paper, we propose Neyman's smooth tests for assessing the normality assumption across various types of ANOVA models. The proposed test statistics are constructed based on the Gaussian probability integral transformation of ANOVA residuals. Under the null hypothesis of normality, the test statistics are asymptotically chi-square distributed, with degrees of freedom determined by the dimension of the smooth model (the number of basis functions). A data-driven selection of the model dimension using a modified Schwarz criterion is also discussed. Simulation studies demonstrate the effectiveness of the proposed method.

Homogeneity pursuit in ranking inference based on pairwise comparison

♦Yuxin Tao,Tracy Ke

Southern University of Science and Technology, Harvard University

The Bradley-Terry-Luce (BTL) model is one of the most celebrated models for ranking inferences based on pairwise comparison data, which associates individuals with latent preference scores and produces ranks. An important question that arises is the uncertainty quantification for ranks. It is natural to think that the ranks for two individuals are untrustworthy if there is only a subtle difference in their preference scores. In this paper, we explore the homogeneity of scores in the BTL model, which assumes that individuals cluster into groups with the same preference scores. We introduce the clustering algorithm in regression via data-driven segmentation (CARDS) penalty into the likelihood function, which can rigorously and effectively separate parameters and uncover group structure. Statistical properties of the two versions of CARDS are analyzed. As a result, we achieve a faster convergence rate and sharper confidence intervals for the maximum likelihood estimation (MLE) of preference scores, providing insight into the power of exploring low-dimensional structure in a high-dimensional setting. We analyze the real data, including NBA basketball ranking and journal ranking data, which demonstrate the improved prediction performance and interpretation ability of the proposed method.

Group 4:Causal Inference & Treatment Effects

Double-Score Gaussian Process Model for Robust Causal Inference in Observational Studies

♦Xinyi Xu,Biqing Yang

Ohio State University, Ohio State University

The propensity score is a widely recognized tool for addressing confounding in causal inference within observational data. Recently, matching methodologies have evolved to incorporate both prognostic and propensity scores, offering consistent estimations when either model is correctly specified. However, these methods have centered on estimating Population Average Treatment Effects (PATE) but overlook the heterogeneity of treatment effects across subgroups, or Conditional Average Treatment Effects (CATE), which are crucial in many practical scenarios. Also, assessing estimation variability for matching estimators can be challenging and computationally demanding.

In this work, we propose a Bayesian semi-parametric causal model that integrates both propensity and prognostic scores through Gaussian processes. This model exhibits a double robustness attribute, ensuring consistency if either score model is accurate. Furthermore, we derive the asymptotic distributions for both PATE and CATE under our model and a propensity-score-only model. This comparison reveals that our double score model offers reduced asymptotic variances in estimating both PATE and CATE. The implementation of a Markov Chain Monte Carlo (MCMC) algorithm within this framework facilitates sampling from posterior distributions, which allows straightforward calculations of the variances for both PATE and CATE estimators. Through extensive simulation studies, we demonstrate the superior performance of our double-score model over the propensity-score-only model, particularly when faced with model misspecification. This advancement represents a significant stride in the field of causal inference, offering more robust and nuanced insights in observational studies.

Deep Orthogonal Learner for Conditional Quantile Treatment Effect Estimation

♦Qixian Zhong

Xiamen University

In recent years, orthogonal statistical learning has been widely recognized for its ability to reduce sensitivity with respect to nuisance parameters to estimate target parameter, making it an important tool in causal inference, particularly in the estimation of the conditional average treatment effect (CATE). However, its application on conditional quantile treatment effect (CQTE), which offers a more expansive view of the treatment effect than

CATE, has not yet been explored comprehensively. In this paper, we propose a novel method for learning CQTE. This method shares Neyman orthogonal property, which produces CQTE estimators that are insensitive to small perturbations of nuisance functions. We first model the CQTE nonparametrically and use deep learning to approach it. We establish the convergence rate of the neural network estimator, demonstrating that it achieves the minimax optimal rate of convergence (up to a polylogarithmic factor). This highlights deep learning's ability to identify low-dimensional structures in high-dimensional data. Additionally, we then model CQTE linearly to facilitate interpretation and statistical inference. We prove that the corresponding coefficient and CQTE estimators achieve root-n consistency and asymptotic normality, even if the estimators of the nuisance parameters converge at a slower rate. Through empirical evaluation for numerical studies, we demonstrate the superiority of our method compared to competing methods.

Efficient Semi-supervised Estimation of Optimal Individualized Treatment Regime with Survival Outcome

♦*Xintong Li, Jialiang Li,Yong Zhou*

East China Normal University, National University of Singapore, East China Normal University

We consider the problem of estimating optimal Individualized Treatment Regime (ITR) for survival outcome under right-censoring in a semi-supervised learning framework. In this setting, only a small subset of observations is labeled with the true outcomes of interest, while the majority remain unlabeled. To address this challenge, we propose an imputation-based semi-supervised approach which is robust and adaptable to all sorts of imputation models. We employ a flexible single-index kernel smoothing imputation technique to effectively utilize the unlabeled data in multidimensional covariate settings. The proposed estimators for the parameters indexing the optimal ITR are shown to be consistent and asymptotically normal. Moreover, the semi-supervised estimation enhances efficiency by reducing asymptotic variance relative to the supervised estimation. Numerical experiments on both simulated and real datasets demonstrate the superior performance of our proposed semi-supervised approach.

Optimal Designs for Order-of-Addition Two-Level Factorial Experiments

♦*Qiang Zhao,Qian Xiao,Abhyuday Mandal,Fasheng Sun*

Northeast Normal University, Shanghai Jiao Tong University, University of Georgia, Northeast Normal University

A new type of experiment, called the order-of-addition factorial experiment, has recently received considerable attention in medical science and bioengineering. These experiments aim to simultaneously optimize the order of addition and dose levels of drug components. In the experimental design literature, the concept of dual-orthogonal arrays (DOAs), a class of optimal order-of-addition two-level factorial designs under the compound model, has recently been introduced for such experiments. However, constructing flexible DOAs remains a challenging task. In this paper, we propose a novel theory-guided search method to efficiently identify DOAs of any feasible size. We also provide an algebraic construction method that immediately leads to certain DOAs. Moreover, to address the potential issue that DOA ignores interaction effects, we propose to construct a new type of optimal designs under the expanded compound model, named the strong DOA (SDOA). We provide two algebraic construction methods for the SDOA. We establish theoretical results on the optimality of both DOAs and SDOAs. Numerical studies illustrate the superiority of our proposed designs.

Group 5: Network & Graph Analysis

Systemic Risk Management via Maximum Independent Set in Extremal Dependence Networks

♦*Qian Hui,Tiandong Wang*

Fudan University, Fudan University

The failure of key financial institutions may accelerate risk contagion due to their interconnections within the system. In this paper, we propose a robust portfolio strategy to mitigate systemic risks during extreme events. We use the stock returns of key financial institutions as an indicator of their performance, apply extreme value theory to assess the extremal dependence among stocks of financial institutions, and construct a network model based on a threshold approach that captures extremal dependence. Our analysis reveals different dependence structures in the Chinese and U.S. financial systems. By applying the maximum independent set (MIS) from graph theory, we identify a subset of institutions with minimal extremal dependence, facilitating the construction of diversified portfolios resilient to risk contagion. We also compare the performance of our proposed portfolios with that of the market portfolios in the two economies.

Testing Global Community Structure in Multi-Layer Networks: A Leave-One-Out Polynomial Statistic

♦*Xiyue Zhu, Xiao Han*

University of Science and Technology of China, University of Science and Technology of China

In many practical scenarios, single-layer networks often fail to provide sufficient signals for accurately reconstructing community structures, emphasizing the need to analyze multi-layer networks. This paper addresses the fundamental problem of determining whether a multi-layer network consists of a single community or multiple communities. We propose a novel multi-layer leave-one-out order-4 polynomial statistic designed to extract and leverage cross-layer information for testing global community structure under the multi-layer mixed membership stochastic block model (MLMSBM). Theoretical analysis establishes the asymptotic null distribution of the proposed statistic and demonstrates its power under mild conditions. Notably, the theoretical condition achieves a nearly optimal network density threshold for community detection among all polynomial-time algorithms. Simulations and real data examples show the effectiveness of our statistic.

Spatially aware adjusted Rand index for evaluating spatial transcriptomics clustering

♦*Yinqiao Yan,Xiangnan Feng,Xiangyu Luo*

Beijing University of Technology, Fudan University, Renmin University of China

The spatial transcriptomics (ST) clustering plays a crucial role in elucidating the tissue spatial heterogeneity. An accurate ST clustering result can greatly benefit downstream biological analyses. As various ST clustering approaches are proposed in recent years, comparing their clustering accuracy becomes important in benchmarking studies. However, the widely used metric, adjusted Rand index (ARI), totally ignores the spatial information in ST data, which prevents ARI from fully evaluating spatial ST clustering methods. We propose a spatially aware Rand index (spRI) as well as spARI that incorporate the spatial distance information. Specifically, when comparing two partitions, spRI provides a disagreement object pair with a weight relying on the distance of the two objects, whereas Rand index assigns a zero weight to it. This spatially aware feature of spRI adaptively differentiates disagreement object pairs based on their distinct distances, providing a useful evaluation metric that favors spatial coherence of clustering. The spARI is obtained by adjusting spRI for random chances such that its expectation takes zero under an appropriate null model. Statistical properties of spRI and spARI are discussed. The applications to simulation study and two ST datasets demonstrate the improved utilities of spARI compared to ARI in evaluating ST clustering methods.

Rate Guarantees for recovery of latent space distances

♦Yunhe Pan, Pavel Krivitsky, Feng Chen

We consider the problem of predicting relationships (the presence or absence of a tie) in a network from

partially observed data. We assume the sampled network is generated from a dyad independent model which is parameterised by some matrix M , such that the probability of forming a link between a pair of actors is a function of the corresponding matrix entry. The dyad Yi,j is a Bernoulli Random Variable with probability of success given by F (Mi,j)

The task is to recover the matrix M from the available observations (observed present, observed absent, or missing) which are generated under this model.If we impose restrictions on the class of matrices that are allowed in the model, then estimation of M becomes tractable. In particular, we discuss the case where the parameters space is the class of distance matrices.

In this situation, low-rank structure may arise from the presence of latent variables, such that the model's intrinsic degrees of freedom are smaller than its extrinsic dimensionality. We discuss as an application, link prediction in latent space models (Hoff 2002) and provide a bound on the estimation error.

Group 6:Time Series & Functional Data

Optimal Subsampling and EM Algorithms for Non-Markovian Semiparametric Regression with Interval-Censored Multi-State Data

♦Si Cheng Fong,Tony Sit, Hoi Ying Wong

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

Interval-censored multi-state data commonly arise in longitudinal studies of chronic diseases and credit rating transitions, where subjects transition among discrete states with exact transition times unobserved—only known to lie within certain intervals. Estimation in such settings presents three major challenges: (1) identifiability issues, particularly when subjects revert to previous states within unobserved intervals; (2) non-Markovian dynamics, where transition intensities depend on latent historical trajectories due to censoring; and (3) computational inefficiency in large-scale datasets.

To address these challenges, we propose a semiparametric proportional intensity model with well-defined identifiability conditions that guarantee unique parameter estimation. We employ flexible B-spline estimators to accommodate non-Markovian effects and develop a stable expectation-maximization (EM) algorithm for nonparametric maximum likelihood estimation under general interval censoring. Furthermore, we introduce a general subsampling EM framework, which we advance into an optimal subsampling EM algorithm by assigning optimal sampling probabilities to censored subjects. This innovation significantly improves computational efficiency of the parameters of interest while achieving optimal efficiency. Theoretical results establish the consistency, asymptotic normality, and semiparametric efficiency of the proposed estimators. Extensive simulations demonstrate the method's robustness, scalability, and practical utility in analyzing complex interval-censored multi-state processes.

A Unified Principal Component Analysis for Stationary Functional Time Series

♦Zerui Guo,Jianbin Tan,Hui Huang

School of Mathematics, Sun Yat-sen University, Duke University, Renmin University of China

Functional time series (FTS) data have become increasingly available in real-world applications. Research on such data typically focuses on two objectives: curve reconstruction and forecasting, both of which require efficient dimension reduction. While functional principal component analysis (FPCA) serves as a standard tool, existing methods often fail to achieve simultaneous parsimony and optimality in dimension reduction, thereby restricting their practical implementation. To address this limitation, we propose a novel notion termed optimal functional filters, which unifies and enhances conventional FPCA methodologies. Specifically, we establish connections among diverse FPCA approaches through a dependence-adaptive representer for stationary FTS. Building on this theoretical foundation, we develop an estimation procedure for optimal functional filters that enables both dimension reduction and prediction within a Bayesian hierarchical modeling framework. Theoretical properties are established for the proposed methodology, and comprehensive simulation studies validate its superiority over competing approaches. We further demonstrate our method through an application to reconstructing and forecasting daily air pollutant concentration trajectories.

From sparse to dense functional time series: phase transitions of detecting structural breaks and beyond

Leheng Cai, ♦Qirui Hu

Tsinghua University, Tsinghua University

We develop a novel methodology for detecting abrupt break points in mean functions of functional time series, adaptable to arbitrary sampling schemes. By employing B-spline smoothing, we introduce $mathcal L_{infty}$ and $mathcal L_2$ test statistics statistics based on a smoothed cumulative summation (CUSUM) process, and derive the corresponding asymptotic distributions under the null and local alternative hypothesis, as well as the phase transition boundary from sparse to dense. We further establish the convergence rate of the proposed break point estimators and conduct statistical inference on the jump magnitude based on the estimated break point, also applicable across sparsely, semi-densely, and densely, observed random functions. Extensive numerical experiments validate the effectiveness of the proposed procedures. To illustrate the practical relevance, we apply the developed methods to analyze electricity price data and temperature data.

Two-way Matrix Autoregressive Model with Thresholds

♦Cheng Yu, Dong Li, Xinyu Zhang, Howell Tong

Tsinghua University, Tsinghua University, University of Iowa, Tsinghua Univeristy and London School of Economics and Political Science

Recently, matrix-valued time series data have attracted significant attention in the literature with the recognition of threshold nonlinearity representing a significant advance. However, given the fact that a matrix is a two-array structure, it is unfortunate, perhaps even unusual, for the threshold literature to focus on using the same threshold variable for the rows and the columns. In fact, evidence in economic, financial, environmental and other data shows advantages of allowing the possibilities of two different threshold variables (with possibly different threshold parameters for rows and columns), hence the need for a Two-way Matrix AutoRegressive model with Thresholds (2-MART). Naturally, two threshold variables pose new and perhaps even fierce challenges, which might be the reason behind the adoption of only one threshold variable in the literature up to now. In this paper, we develop a comprehensive methodology for the 2-MART model, by overcoming various challenges. Compared with existing models in the literature, the new model can achieve greater dimension reduction, much better model fitting, more accurate predictions, and more plausible interpretations.

Group 7: Biostatistics & Medical Applications

Mixed membership latent variable model with unknown factors, factor loadings and number of extreme profiles

♦Yuyang He, Kai Kang, Xinyuan Song

The Chinese University of Hong Kong, Sun Yat-sen University, The Chinese University of Hong Kong

Mixed membership models are frequently utilized to capture complex individual heterogeneity in multivariate and longitudinal data. A key aspect of mixed membership modeling involves determining the number of extreme profiles (classes), a task traditionally managed through inefficient criterion-based methods. This task is particularly challenging when the predictors within the models are latent and derived from multiple observed variables using exploratory factor analysis. In this paper, we consider an innovative mixed membership latent variable model, which consists of an exploratory factor model to identify latent factors and a mixed membership model with latent predictors. We develop an efficient approach that integrates parameter estimation and model selection for the number of factors, extreme profiles, and the structure of the factor loading matrix. Our approach comprises a modified stochastic search item selection algorithm to automatically determine the number of latent factors and their associated manifest variables and a Bayesian penalized method to select the number of extreme profiles. We validate our methodology through extensive simulation studies, demonstrating its accuracy and efficiency in both parameter estimation and model selection. Applying this method to data from the Parkinson's Progression Markers Initiative, we identify clinically important latent traits and distinct disease profiles. The results underscore our model's enhanced ability to depict the intricate individual heterogeneity present in Parkinson's disease (PD) patients.

Joint Modeling Approach for censored predictors in generalized linear model due to detection limit with applications to metabolites data

Fengxue Li,Qinglin Wang,Wan Tang,Peng Ye,♦Hua He

Tulane University, University of International Business and Economics, Tulane University, University of International Business and Economics, Tulane University

Biomarker measurements obtained from urine, serum, or other biological matrices are commonly utilized in medical and health-related research. However, it is frequently observed that biomarker measures are left-censored due to their concentration levels falling below the limits of detection of the assay. When biomarker measurements are left-censored, they can originate from non-exposed subjects where their measures consistently register as zeros, thus being censored, or from exposed subjects whose exposure levels are below the detection limit threshold. In cases where censored biomarkers originate from both exposed and non-exposed subjects, the study population becomes mixed, valid inference is only assured if the mixed population is disentangled in data analysis. In our paper, we extend the joint modeling approach to handle censored predictors in generalized linear models. Intensive simulation studies are conducted to assess the performance of the joint modeling approach, and the results demonstrate that these approaches can provide unbiased and efficient estimates. Additionally, we apply the joint modeling approaches to examine associations between plasma metabolites and hypertension in the Bogalusa Heart Study, identifying new metabolites highly associated with hypertension

Double Optimal Transport for Differential Gene Regulatory Network Inference with Unpaired Samples

♦Mengyu Li,Bencong Zhu,Cheng Meng,Xiaodan Fan, ,

Renmin University of China,The Chinese University of Hong Kong, Renmin University of China, The Chinese University of Hong Kong

Inferring differential gene regulatory networks (GRNs) between different conditions from gene expression profiles remains a significant challenge. Current GRN inference approaches are limited by either scalability in large networks or accuracy in high-dimensional scenarios. Furthermore, most existing methods require paired samples for comparative GRN analyses. To overcome these challenges, we model gene regulation as a distribution transportation problem and propose an efficient and effective method, called Double Optimal Transport (OT), for reconstructing differential GRNs from the perspective of optimal transport theory, applicable to unpaired samples. Double OT is a novel two-level OT framework. It first aligns unpaired samples by solving a partial OT problem at the sample level, and then infers GRNs from the aligned samples by solving a robust OT problem at the gene level. Comprehensive simulation studies demonstrate the superior efficiency and efficacy of Double OT in different scales of networks compared to state-of-the-art methods. We also apply the proposed method to a gastric cancer dataset, identifying the proto-oncogene MET as a central node in the gastric cancer GRN. Its crucial role in early oncogenesis and potential as a therapeutic target further validate our approach and enhance our understanding of the regulatory mechanisms of gastric cancer.

Nested Deep Learning Model Towards a Foundation Model for Brain Signal Data

♦*Fangyi Wei, Jiajie Mo, Kai Zhang, Haipeng Shen, Srikantan Nagarajan, Fei Jiang*

University of Hong Kong, Beijing Tiantan Hospital, Beijing Tiantan Hospital, The University of Hong Kong, University of California, University of California

Epilepsy affects around 50 million people globally. Electroencephalography (EEG) or Magnetoencephalography (MEG) based spike detection plays a crucial role in diagnosis and treatment. Manual spike identification is time-consuming and requires specialized training that further limits the number of qualified professionals. To ease the difficulty, various algorithmic approaches have been developed. However, the existing methods face challenges in handling varying channel configurations and in identifying the specific channels where the spikes originate. A novel Nested Deep Learning (NDL) framework is proposed to overcome these limitations. NDL applies a weighted combination of signals across all channels, ensuring adaptability to different channel setups, and allows clinicians to identify key channels more accurately. Through theoretical analysis and empirical validation on real EEG/MEG datasets, NDL is shown to improve prediction accuracy, achieve channel localization, support cross-modality data integration, and adapt to various neurophysiological applications.

Group 8: Dimension Reduction & Variable Selection

False Discovery Control for High-Dimensional Linear Models with Model-X Knockoff and p-values

Jinyuan Chang, Chenlong Li, Cheng Yong Tang, ♦*Zhengtian Zhu*

Southwestern University of Finance and Economics, School of Mathematics, Taiyuan University of Technology, Temple University, Chinese Academy of Sciences

In this study, we propose new false discovery rate (FDR) control methods using p-values evaluated from carefully constructed test statistics to address the multiple testing problem in high-dimensional linear models. We tackle two primary challenges: the inherently daunting task of statistical inference in high-dimensional models and the methodological and theoretical difficulties in developing appropriate approaches that ensure valid FDR control under complex and unknown dependencies among test statistics. To overcome these challenges, we develop a novel multiple testing framework, devising model-X knockoff variables, that integrates debiasing techniques with a penalized regression estimator applied to high-dimensional linear models. The framework utilizes an augmented model matrix containing both the original variables and their knockoff counterparts. Through an appropriate linear transformation upon debiasing the penalized regression estimator, we construct naturally paired statistics and their associated p-values for high-dimensional hypotheses. Based on these paired p-values, we implement a two-step multiple testing procedure: the first step applies the Bonferroni method for initial selection, followed by the Benjamini-Hochberg procedure to finalize decisions – each using one of the two sets of paired p-values. We establish the theoretical validity of the debiasing process and confirm that the proposed methods effectively control the FDR. Empirical studies demonstrate that our approach outperforms competing techniques based on empirical false discovery proportions, achieving valid FDR control and enhanced statistical power, particularly in scenarios with weaker signals, smaller sample sizes, and lower FDR levels.

Multi-Population Sufficient Dimension Reduction

♦*Meggie Wen, Yuexiao Dong, Li-Xing Zhu*

Missouri University of Science and Technology, Temple University, Beijing Normal University at Zhuhai

In this paper, we propose a novel dimension reduction method for multi-population data. The method conducts a joint analysis that utilizes all information between populations while still retaining group-specific effects.Contrary to partial dimension reduction methods primarily identifying related directions across all populations and the conditional analysis conducted independently within each individual population, our two-step method makes use of the information across the multiple populations which improves the estimation accuracy. Simulations and a real data example are given to illustrate our methodology.

SOFARI-R: High-Dimensional Manifold-Based Inference for Latent Responses

Zemin Zheng, ♦*Xin Zhou, Jinchi Lv*

University of Science and Technology of China, University of Science and Technology of China, University of Southern California

Data reduction with uncertainty quantification plays a key role in various multi-task learning applications, where large numbers of responses and features are present. To this end, a general framework of high-dimensional manifold-based SOFAR inference (SOFARI) was introduced recently in Zheng, Zhou, Fan and Lv (2024) for interpretable multi-task learning inference focusing on the left factor vectors and singular values exploiting the latent singular value decomposition (SVD) structure. Yet, designing a valid inference procedure on the latent right factor vectors is not straightforward from that of the left ones and can be even more challenging due to asymmetry of left and right singular vectors in the response matrix. To tackle these issues, in this paper we suggest a new method of high-dimensional manifold-based SOFAR inference for latent responses (SOFARI-R), where two variants of SOFARI-R are introduced. The first variant deals with strongly orthogonal factors by coupling left singular vectors with the design matrix and then appropriately rescaling them to generate new Stiefel manifolds. The second variant handles the more general weakly orthogonal factors by employing the hard-thresholded SOFARI estimates and delicately incorporating approximation errors into the distribution. Both variants produce bias-corrected estimators for the latent right factor vectors that enjoy asymptotically normal distributions with justified asymptotic variance estimates. We demonstrate the effectiveness of the newly suggested method using extensive simulation studies and an economic application.

Overview of normal-reference tests for high-dimensional means with implementation in the R package 'HDNRA'

♦*Pengfei Wang, Tianming Zhu, Jin-Ting Zhang*

Nanyang Technological University, Nanyang Technological University, National University of Singapore

The challenge of testing for equal mean vectors in high-dimensional data poses significant difficulties in statistical inference. Much of the existing literature introduces methods that often rely on stringent regularity conditions for the underlying covariance matrices, enabling asymptotic normality of test statistics. However, this can lead to complications in controlling test size. To address these issues, a new set of tests has emerged, leveraging the normal-reference approach to improve reliability. The latest normal-reference methods for testing equality of mean vectors in high-dimensional samples, potentially with differing covariance structures, are reviewed. The theoretical underpinnings of these tests are revisited, providing a new unified justification for the validity of centralized L^2-norm-based normal-reference tests (NRTs) by deriving the convergence rate of the distance between the null distribution of the test statistic and its corresponding normal-reference distribution. To facilitate practical application, an R package, HDNRA, is introduced, implementing these NRTs and extending beyond the two-sample problem to accommodate general linear hypothesis testing (GLHT). The package, designed with user-friendliness in mind, achieves efficient computation through a core implemented in C++ using Rcpp, OpenMP, and RcppArmadillo. Examples with real datasets are included, showcasing the application of various tests and providing insights into their practical utility.

www.icsa.org

International Chinese Statistical Association

泛華統計協會