浙江大学 数据科学研究中心
Center for Data Science
ZHEJIANG UNIVERSITY

2026

International Conference on Frontiers of Data Science

会议手册

Program

Photo Live Streaming

May 15-17, 2026

Hangzhou·China

Center for Data Science, Zhejiang University

Contents

About Zhejiang University.....1

About Center for Data Science.....3

Conference Committees.....5

Transportation.....6

Venue Map.....7

Dinning.....9

Program Overview.....10

Scientific Program.....14

Keynote Speech.....27

Abstracts

2026-05-16 10:35-11:55

Dongwan Hall A: Machine Learning for Policy Optimization and Causal Inference in Complex Systems

Speakers: Qingliang Fan, Yunan Wu, Xiangnan Feng, Haoran Xue.....29

Dongwan Hall B: Advances in Causal Inference for Observation Data

Speakers: Hongzhe Li, Yumou Qiu, Wang Miao, Ziwei Mei.....31

Meishu Hall: Statistical and Machine Learning Methods for Analyzing Complex Real-world Data

Speakers: Yiqun Chen, Hao Mei, Xingjie Shi, Hang Zhou.....33

Meizhu Hall: Advances in Machine Learning Theory

Speakers: Yuan Cao, Yunwen Lei, Gen Li, Wei Huang.....35

Jiaolu Hall A: Statistical Theory and Random Matrices with Applications in Finance

Speakers: Xinghua Zheng, Henry Reeve, Yukun He.....37

Jiaolu Hall B: Deep Generative Models and Their Interplay with Data-driven Decision-making

Speakers: Jian Qian, Zhun Deng, Nian Si, Yuchen Zhou.....39

Yanshui Hall: From Statistical Foundations to AI Impacts: Risk, Robustness, and Representation

Speakers: Xiangchao Li, Yanlin Hu, Lingchong Liu, John Park.....42

Shangwu Hall: Statistical Learning for Complex Systems

Speakers: Shang Wu, Wei Zhang, Guorong Dai, Chengli Tan.....44

2026-05-16 14:00-15:20

Dongwan Hall A: Recent Advances in Statistics and AI
Speakers: Zhanrui Cai, Ruijia Wu, Sai Li, Shuting Shen……46
Dongwan Hall B: Recent Developments in Hypothesis Testing
Speakers: Yin Xia, Yeqing Zhou, Shiyang Ma, Lucy Xia……48
Meishu Hall: Recent Advances in Estimation and Inference for Structured and Heterogeneous Data
Speakers: Liujun Chen ,Zhixiang Zhang, Yaoming Zhen……50
Meizhu Hall: Prediction and Machine Learning in Biomedical Studies
Speakers: Guogen Shan, Xiaoyu Song, Kam Lun Tsang, Quran Wu……52
Jiaolu Hall A: Private and Federated High-Dimensional Statistical Inference
Speakers: Xin Chen, Lin Yang, Dong Xia……55
Jiaolu Hall B: Transfer Learning and Robust Modeling for Structured Data
Speakers: Yingying Li, Xuejun Jiang, Zirui Wang, Hao Zeng……57
Yanshui Hall: Advanced Data Science Techniques: Differential Privacy, Dynamic Treatment Regimes
Speakers: Shuyuan Wu, Yicheng Li, Haobo Qi, Yuqian Zhang……59
Shangwu Hall: Recent Advances in High-dimensional and Financial Statistics
Speakers: Jingkun Qiu, Wu Su, Yudong Wang, Yushan Xue……61

2026-05-16 15:50-17:10

Dongwan Hall A: Optimal Inference under Privacy, Communication, and Sampling Constraints
Speakers: Mengchu Li, Linjun Zhang, Lasse Vuursteen……63
Dongwan Hall B: Advances in High-Dimensional Inference and Uncertainty Quantification
Speakers: Jingyuan Liu，Zijian Guo，Juan Shen，Haojie Ren……65
Meishu Hall: Regression and Time-series in High-dimensional Setting
Speakers: Shanshan Song, Qianqian Zhu, Huiling Yuan, Lan Li……67
Meizhu Hall: Robust and Efficient Estimation for Modern Causal and Treatment Effect Models
Speakers: Wei Huang, Mengchu Zheng, Lin Liu……69
Jiaolu Hall A: Advances in Statistical Methods for Multiple Biomarkers and Data Sources
Speakers: Jialiang Li，Wenlong Mou，Jiarui Zhang，Yao Zhang……71
Jiaolu Hall B: Recent Advances in Statistical Genetics
Speakers: Zilin Li, Yaowu Liu, Xianghong Hu, Jingsi Ming……73
Yanshui Hall: New Advances in Data Science
Speakers: Tao Wang，Zhonghua Liu，Jiang Gui，Keren Li……75
Shangwu Hall: Multiple Testing Meets Modern Data and Machine Learning
Speakers: Xiaoyang Wu, Yuan Yao, Bowen Gang, Zinan Zhao……77

2026-05-17 10:20-11:40

Dongwan Hall A: Model Averaging and Related Studies (MARS)
Speakers: Xiangyu Cui, Fang Fang, Dandan Jiang, Dalei Yu……79

Dongwan Hall B: Statistical Learning on Complex Data Analysis
Speakers: Zhenhua Lin, Wenliang Pan, Long Feng, Ting Li……...……...……..81
Meishu Hall: Statistical Learning Theory and Methods: From High-Dimensional Inference to Multi-Modal AI Applications
Speakers: Wei Liu, Pengkun Yang, Lican Kang, Ling Zhou…….83
Meizhu Hall: Recent Advances in Statistical Theory for Deep Learning
Speakers: Yuling Jiao, Huiming Zhang, Guohao Shen, Juntong Chen…….85
Jiaolu Hall A: Frontiers in Interactive Generative World Models
Speakers: Lu Sheng，Yudong Guo，Zhiwen Shao，Keyu Chen……..87
Jiaolu Hall B: Statistical Analysis of Network
Speakers: Yan Zhang, Yongqin Qiu, Yi Ding, Yuyang Liu……..89
Yanshui Hall: Modern Methods in Statistical Learning and Data Analysis
Speakers: Ruijian Han, Yu Gu, Chendi Wang, Yuxin Tao…….……..91

2026-05-17 14:00-15:20
Dongwan Hall A: Recent Advances in the Statistical Foundations of Data Sciences
Speakers: Anderson Ye Zhang, Zhengyu Huang, Weichen Wang, Feiyu Jiang……...…93
Dongwan Hall B: Statistical Inference and Sequential Decision Making
Speakers: Yiyuan She, Bangyao Zhao, Ran Chen, Kaijie Xue……..…95
Meishu Hall: Deep Learning Methods in Survival Analysis
Speakers: Xingqiu Zhao, Wen Yu, Qixian Zhong, Zhangsheng Yu……..…97
Meizhu Hall: Recent Advances in Survival Analysis
Speakers: Yongfu Yu, Fangyao Chen, Huijuan Ma, Chengfeng Zhang……...99
Jiaolu Hall A: Advances in Intelligent Agents and Efficient Multimodal Models
Speakers: Chi Zhang, Wangbo Zhao, Xu Yang, Bohan Zhuang…….101
Jiaolu Hall B: Preference-based centrality and ranking in general metrics space
Speakers: Doudou Zhou, Mengyan Li, Yuming Zhang, Alexander Giessing…….103
Yanshui Hall: Robust Treatment Effect Learning and Fair Decision Making
Speakers: Changcheng Li, Molei Liu, Zeyu Bian, Ping Zhang…….105

2026-05-17 15:50-17:10
Dongwan Hall A: Statistical Learning: Theory and Practice
Speakers: Yuheng Ma, Jing Zeng, Rui Qiu, Ziwen Gao……..…107
Dongwan Hall B: Machine Learning and Statistical Inference for Complex Data
Speakers: Jianqiao Wang, Xin Cong, Huaqing Jin, Hanzhong Liu…….…...109
Meishu Hall: Next Generation Visual AIGC: Controllable Generation, Ultra-HD, and Intelligent Design
Speakers: Xiaodong Cun, Qian Yu, Ying Tai, Chunxia Xiao…….111
Meizhu Hall: Advances in Statistical Learning and High-Dimensional Statistical Inference
Speakers: Jun Shu, Danning Li, Baihua He, Zhengtian Zhu……...…113

Jiaolu Hall A: Recent Advances in Causal Inference and Latent Variable Modeling

Speakers: Xin Zhang, Zheng Zhang, Ruiyi Yang, Yujia Gu……...……115

Jiaolu Hall B: Advances in Inference for Modern Data Settings

Speakers: Stéphane Guerrier, Yanyuan Ma, Mucyo Karemera, Luca Insolia……..…..117

Yanshui Hall: Methods in Spatial Transcriptomic Data Analysis

Speakers: Chunman Zuo, Hongyi Xin, Xin Yuan, Ruitian Gao……...…..……...…..119

浙江大学概况

浙江大学是一所历史悠久、声誉卓著的高等学府，坐落于中国历史文化名城、风景旅游胜地杭州。浙江大学的前身求是书院创立于 1897 年，为中国人自己最早创办的新式高等学校之一。1928 年，定名国立浙江大学。抗战期间，浙大举校西迁，在贵州遵义、湄潭等地办学七年，1946 年秋回迁杭州。1952 年全国高等学校院系调整时，浙江大学部分系科转入兄弟高校和中国科学院，留在杭州的主体部分被分为多所单科性院校，后分别发展为原浙江大学、杭州大学、浙江农业大学和浙江医科大学。1998 年，同根同源的四校实现合并，组建了新浙江大学，迈上了创建世界一流大学的新征程。在 128 年的办学历程中，浙江大学始终秉承以“求是创新”为校训的优良传统，逐步形成了“勤学、修德、明辨、笃实”的浙大人共同价值观和“海纳江河、启真厚德、开物前民、树我邦国”的浙大精神。

浙江大学是一所特色鲜明、在海内外有较大影响的综合型、研究型、创新型大学，学科涵盖哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、管理学、艺术学、交叉学科等 13 个门类，设有 7 个学部、40 个专业学院（系）、1 个工程师学院、2 个中外合作办学学院、7 家直属附属医院。学校现有紫金港、玉泉、西溪、华家池、之江、舟山、海宁等 7 个校区，占地面积 7430707 平方米，图书馆总藏书量 819.2 万册。截至 2024 年底，学校有全日制学生 70035 人、国际学生 6360 人、教职工 9649 人。2022 年，浙江大学入选第二轮“双一流”建设高校，21 个学科入选一流学科建设名单，绝大多数学科在第五轮学科评估中取得可喜进步。

浙江大学紧紧围绕“德才兼备、全面发展”的核心要求，全面落实立德树人根本任务，加快构建以学生成长为中心的卓越教育体系，着力培养德智体美劳全面发展、具有全球竞争力的高素质创新人才和领导者。在长期的办学历程中，学校涌现出大批著名科学家、文化大师以及各行各业的精英翘楚，包括 1 位诺贝尔奖获得者、5 位国家最高科技奖得主、4 位“两弹一星”功勋奖章获得者、1 位“八一勋章”获得者、1 位全军挂像英模、5 位国家荣誉称号获得者、6 位“最美奋斗者”和 230 余位两院院士等杰出典型，为实现中华民族伟大复兴、推进人类文明交流互鉴作出了积极贡献。

浙江大学注重精研学术和科技创新，主动服务重大战略需求，加快打造国家战略科技力量，建设了一批开放性、国际化的高端学术平台，汇聚了各学科的学者大师和高水平研究团队，产出了以国家科技进步特等奖为代表的一系列重大科技成果。

“国有成均，在浙之滨”。浙江大学将坚定不移以习近平新时代中国特色社会主义思想为指导，坚持“更高质量、更加卓越、更受尊敬、更有梦想”的战略导向，致力于思想引领和知识创新，培育担当民族复兴大任的时代新人，为中国式现代化和人类文明进步作出卓越贡献。

About Zhejiang University

Zhejiang University is a prestigious institution of higher education with a long history, located in Hangzhou, a renowned historical and cultural city as well as a scenic tourist destination in China. The forerunner of Zhejiang University, Qiushi Academy, was founded in 1897 as one of China's earliest modern-day institutions of higher education independently established by Chinese scholars. Throughout its 128-year history, Zhejiang University has consistently upheld the fine tradition embodied in its motto: Seeking Truth and Pursuing Innovation.

Today, Zhejiang University has developed into a comprehensive, research-intensive, and innovative institution with distinguished characteristics and significant global influence. Its academic disciplines span 13 categories, including Philosophy, Economics, Law, Education, Literature, History, Science, Engineering, Agriculture, Medicine, Management, Arts, and Interdisciplinary Studies, comprising 7 faculties, 40 schools/departments, 1 polytechnic institute, 2 Sino-foreign cooperative education colleges, and 7 affiliated hospitals. In 2022, ZJU was selected into the second round of the national "Double First-Class" initiative. 21 of its disciplines were listed for development as world-class disciplines, and the vast majority of its disciplines showed notable improvement in the fifth round of China's national discipline evaluation. The university operates across 7 campuses spanning 7,430,707 square meters, with a library collection of 8.192 million volumes. As of the end of 2024, the university had a total of 70,035 full-time students, 6,360 international students, and 9,649 faculty members.

Zhejiang University is dedicated to academic excellence and technological innovation, proactively addressing major strategic needs. The university is accelerating its efforts to become a strategic force in science and technology, having established a series of open and internationally-oriented high-end academic platforms. These platforms bring together distinguished scholars and top-tier research teams across disciplines, yielding significant scientific achievements including the prestigious State Preeminence in Science and Technology Award.

Zhejiang University will adhere to its strategic approach of pursuing higher quality, greater excellence, deeper respect, and stronger aspirations. Committed to intellectual leadership and knowledge innovation, the university dedicates itself to cultivating a new generation capable of shouldering the mission of national rejuvenation, thereby making exceptional contributions to the Chinese path to modernization and the advancement of human civilization.

浙江大学数据科学研究中心简介

“浙江大学数据科学研究中心”成立于 2017 年 5 月 18 日，是以统计学、应用数学、计算机科学和管理学为核心支撑学科，以大数据理论、应用研究和人才培养为主的校设学术创新研究机构。研究中心强调与经济学、医学、生命科学、社会学、工学等众多学科领域的交叉融合，在获取基础研究的硕果同时，注重于科研技术成果的转化。

研究中心的核心目标是通过机制创新，汇聚海内外高层次统计学、计算机科学、应用数学、管理科学等数据科学相关人才，组建高水平专业化研究队伍，聚焦大数据分析，机器学习，人工智能等重点方向，开展具有前瞻性的国际领先水平的研究。研究中心现有专任教师 11 人，师资队伍实力雄厚，其中 3 人入选国家高层次人才计划，4 人入选国家青年人才计划，1 人入选浙江省高层次人才计划，1 人入选浙江省青年人才计划。

研究中心设立数据科学教学平台，面向国家战略与产业需求，构建了本硕博多层次培养体系，着重培养复合型数据科学人才。在本科层面，中心开设“数据科学与大数据技术”微专业项目；在硕士层面，开设“数据科学与工程”全日制专业学位项目；同时培养博士研究生，博士生专业涵盖概率论与数理统计、应用数学、计算机科学与技术、管理科学与工程。现有硕博研究生近 110 人。研究中心为学生提供一流的学术环境与发展平台，定期邀请海内外顶尖学者开设前沿学术报告，并资助丰富的海外交流机会，致力于培养掌握尖端科学知识的高水平、精技术数据科学稀缺人才。

在科学研究方面，研究中心聚焦数据科学前沿，在统计决策理论、大规模统计推断、时空数据分析、因果机器学习、数据融合、数据降维、数据治理与服务、精准医学等方向从事学术研究。中心师生研究成果丰硕，在统计学、医学、信息科学、机器学习等领域的顶级国际期刊、计算机顶级会议以及综合类权威期刊上发表了一系列高水平论文，取得了广泛的国际影响力。多名教师担任国内外重要学术期刊编委及专业学会重要职务。

展望未来，研究中心将继续秉承浙江大学“求是创新”校训，充分发挥杭州大数据产业集聚优势和浙江大学多学科交叉融合特色，聚焦解决重大现实问题，致力于创造具有实际价值的创新算法，切实攻克中国大数据领域的前瞻性挑战，助力浙江大学早日成为国际数据科学研究领域的重要引领者之一。

About Center for Data Science

The Center for Data Science at Zhejiang University was founded on May 18, 2017. It is a university-established institute dedicated to pioneering research and talent training on data science, with statistics, applied mathematics, computer science, and management science as the foundational disciplines. Rooted in significant progress in fundamental research, the center also focuses on interdisciplinary collaboration with economics, medicine, life science, social science, and engineering, etc., thereby transferring scientific achievements into technological developments that benefit the real world.

The center currently has 11 full-time faculty members, including three faculty members selected for the national high-level talent program, four faculty members selected for national young talent programs, one faculty member selected for Zhejiang Province high-level talent programs, and one faculty member selected for the Zhejiang Province young talent program. Several faculty members serve as editors of leading international academic journals and hold important positions in professional societies. Their research areas cover statistical decision theory, large-scale statistical inference, spatial-temporal data analysis, causal inference, machine learning theory, data fusion, data dimensionality reduction, data governance and services and precision medicine.

The center has established a multi-level data science education platform. At the undergraduate level, the center offers a "Data Science and Big Data Technology" micro specialty program, opening to students from various majors who are interested in learning data science. At the graduate level, it provides a full-time master’s degree program in "international Master of Data Science", as well as doctoral training in probability theory, statistics, applied mathematics, computer science, and management science. Currently, nearly 110 graduate students are enrolled in these programs. The center provides students with a first-class academic environment and development platform, regularly inviting top international scholars to deliver cutting-edge lectures and offering extensive support for overseas exchange opportunities. Students have published numerous high-impact papers in top-tier international journals in statistics, medicine, information science, and machine learning, as well as premier computer science conferences.

Based on the strengths of Zhejiang University in interdisciplinary cooperation and the vast development of big data industry in Hangzhou, the Center for Data Science at ZJU will continue to tackle major scientific challenges and create innovative solutions to real-world problems, evolving toward a globally leading institution in data science research and application.

Conference Committees

Advisory Committee

Jianfei Cai	Monash University
Tianxi Cai	Harvard University
Tony Cai	University of Pennsylvania
Lu Tian	Stanford University(Chair)
Yazhen Wang	University of Wisconsin-Madison
Ming Yuan	Columbia University
Heping Zhang	Yale University

Local Organizing Committee

Jia Gu	Zhejiang University
Zijian Guo	Zhejiang University(Chair)
Wenguang Sun	Zhejiang University
Xintao Xia	Zhejiang University

Organizer

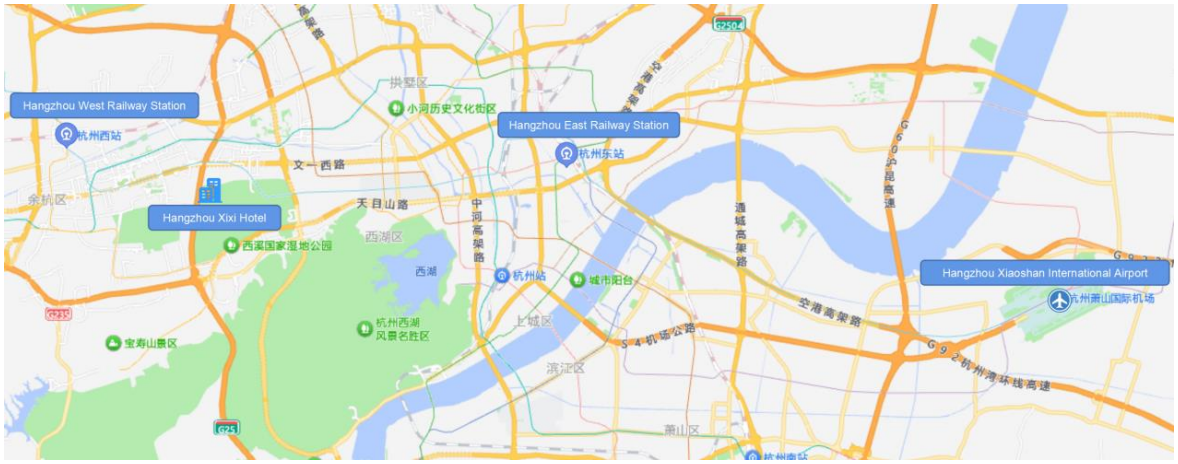
Center for Data Science, Zhejiang University

Official Website

https://ICFDS2026.scimeeting.cn

Transportation

Hangzhou Xixi Hotel can be reached by taxi or public transport from Hangzhou East Railway Station, Hangzhou West Railway Station and Hangzhou Xiaoshan International Airport. Details are as follows:



Transportation options:

Line one:

Hangzhou East Railway Station — Hangzhou Xixi Hotel

Subway Line 19 → 297M | 18km | 51minutes

18 km | 43 minutes | ￥66

Line two:

Hangzhou West Railway Station — Hangzhou Xixi Hotel

Subway Line 19 → 297M | 11 km | 54minutes

11 km | 19minutes | ￥40

Line three:

Hangzhou Xiaoshan International Airport — Hangzhou Xixi Hotel

Subway Line 19 → 297M | 43 kilometers | 1 hour 25 minutes

43 km | 1 hour 2 minutes | ￥90

Venue Map

Hangzhou Xixi Hotel

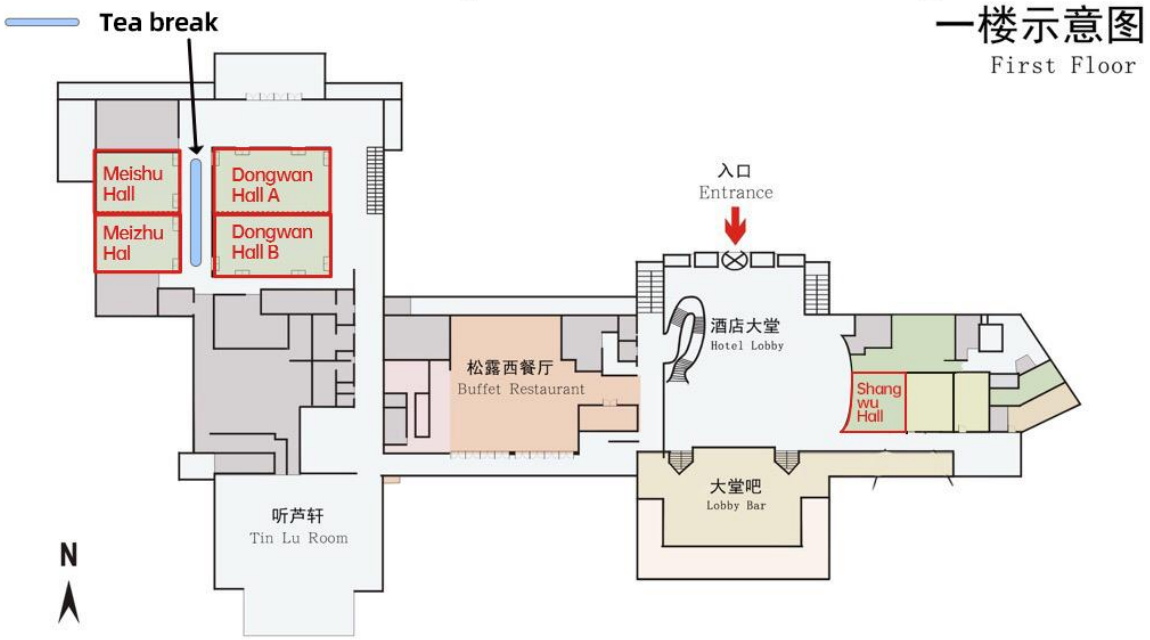
（No.803 Wener West Road，Xihu District）

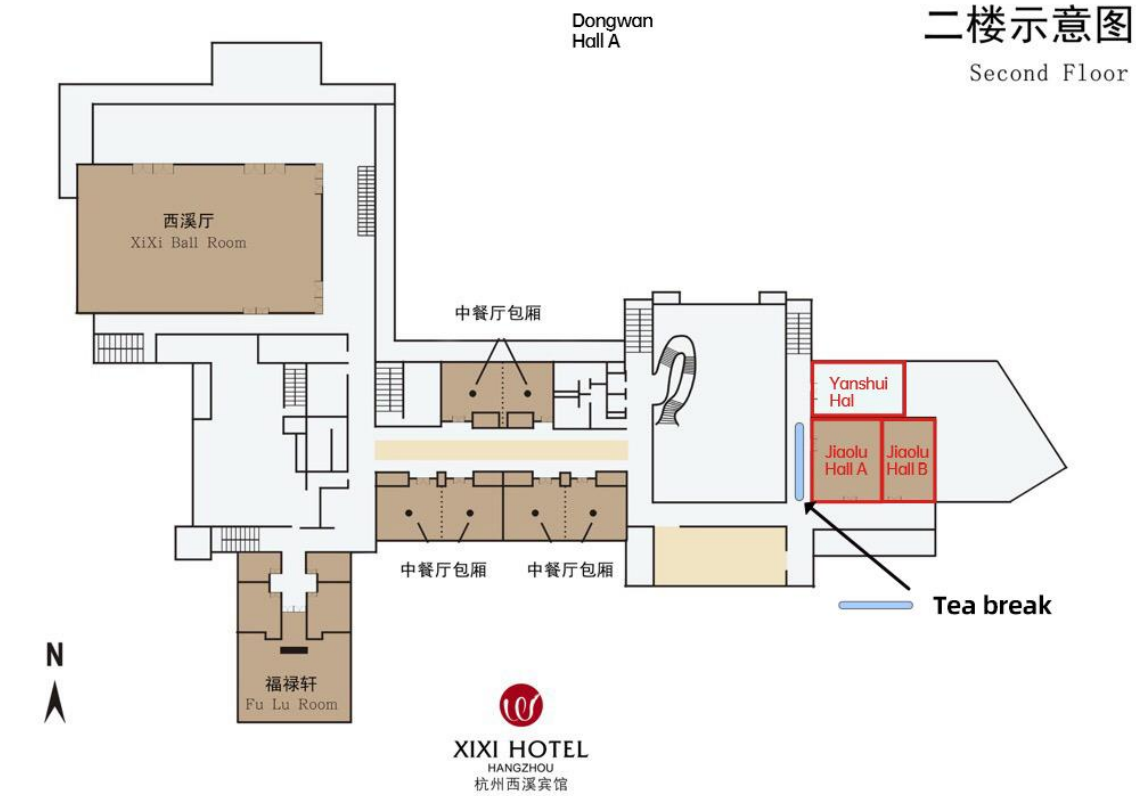
Located at northern part of Xixi Wetland, Hangzhou Xixi Hotel is a 30-minute cab drive from Hangzhou Railway Station and a 1-hour drive from Hangzhou Xiaoshan International Airport. Free Wi-Fi is available within the entire hotel.

Hangzhou Xixi Hotel All units here are equipped with an air conditioning, a flat-screen cable TV and an in-room safe. A mini bar, bottled water and an electric kettle are available. Guests can enjoy the stunning scenery from the window. Hangzhou Xixi Hotel Free toiletries, slippers and hair dryer are found in attached bathroom.A tour desk can arrange a sightseeing tour or rent a car for guests. Currency exchange, laundry and baggage storage are provided at 24-hour front desk.

主楼平面图

General Layout of Main Building





Dinning

Date	Time	Meal	Venue
May 15	17:30-21:00	Dinner	Buffet Restaurant（Truffle） 松露西餐厅 lst floor of the main building
May 16	11:55-14:00	Lunch	Buffet Restaurant（Truffle） 松露西餐厅 lst floor of the main building
May 16	18:00-	Banquet	Dongwan Hall 董湾厅 lst floor of the main building
May 17	11:40-14:00	Lunch	Buffet Restaurant（Truffle） 松露西餐厅 lst floor of the main building

Program Overview

May 16th, Saturday								
09:00-09:15	Opening Ceremony			Chair: Zijian Guo		Dongwan Hall (董湾厅)		
09:15-10:05	Keynote Speech Jun Liu (Tsinghua University)	Title: Conditional sampling via diffusion flow and SMC		Chair: Tony Cai		Dongwan Hall (董湾厅)		
10:05-10:35	**Tea Break**							
10:35-11:55	**Machine Learning for Policy Optimization and Causal Inference in Complex Systems**	**Advances in Causal Inference for Observation Data**	**Statistical and Machine Learning Methods for Analyzing Complex Real-world Data**	**Advances in Machine Learning Theory**	**Statistical Theory and Random Matrices with Applications in Finance**	**Deep Generative Models and Their Interplay with Data-driven Decision-making**	**From Statistical Foundations to AI Impacts: Risk, Robustness, and Representation**	**Statistical Learning for Complex Systems**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)	Shangwu Hall (商务厅)
	Chair:Wei Luo	Chair:Zijian Guo	Chair:Yiqun Chen	Chair:Jun Fan	Chair:Xintao Xia	Chair:Wenlong Mou	Chair:John Park	Chair:Shang Wu
	Organizer: Zijian Guo	Organizer: Zijian Guo	Organizer: Rong Ma	Organizer: Jun Fan	Organizer: Zijian Guo	Organizer: Wenlong Mou	Organizer: Xin Tong	Organizer: Shang Wu
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Qingliang Fan	Hongzhe Li	Yiqun Chen	Yuan Cao	Xinghua Zheng	Jian Qian	Xiangchao Li	Shang Wu
	Yunan Wu	Yumou Qiu	Hao Mei	Yunwen Lei	Henry Reeve	Zhun Deng	Yanlin Hu	Wei Zhang
	Xiangnan Feng	Wang Miao	Xingjie Shi	Gen Li	Yukun He	Nian Si	Lingchong Liu	Guorong Dai
	Haoran Xue	Ziwei Mei	Hang Zhou	Wei Huang		Yuchen Zhou	John Park	Chengli Tan
11:55-14:00	**Lunch** Buffet Restaurant（Truffle）松露西餐厅							
14:00-15:20	**Recent Advances in Statistics and AI**	**Recent Developments in Hypothesis Testing**	**Recent Advances in Estimation and Inference for Structured and Heterogeneous Data**	**Prediction and Machine Learning in Biomedical Studies**	**Private and Federated High-Dimensional Statistical Inference**	**Transfer Learning and Robust Modeling for Structured Data**	**Advanced Data Science Techniques: Differential Privacy, Dynamic Treatment Regimes**	**Recent Advances in High-dimensional and Financial Statistics**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)	Shangwu Hall (商务厅)
	Chair:Linjun Zhang	Chair:Zijian Guo	Chair:Zhongyuan Lyu	Chair:Jiang Gui	Chair:Dong Xia	Chair: Xintao Xia	Chair: Jia Gu	Chair: Yumou Qiu
	Organizer: Linjun Zhang	Organizer: Zijian Guo	Organizer: Zhongyuan Lyu	Organizer: Zhigang Li	Organizer: Dong Xia	Organizer: Zijian Guo&Xintao Xia	Organizer: Jia Gu	Organizer: Yumou Qiu
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Zhanrui Cai	Yin Xia	Liujun Chen	Guogen Shan	Xin Chen	Yingying Li	Shuyuan Wu	Jingkun Qiu
	Ruijia Wu	Yeqing Zhou	Zhixiang Zhang	Xiaoyu Song	Lin Yang	Xuejun Jiang	Yicheng Li	Wu Su
	Sai Li	Shiyang Ma	Yaoming Zhen	Kam Lun Tsang	Dong Xia	Zirui Wang	Haobo Qi	Yudong Wang
	Shuting Shen	Lucy Xia		Quran Wu		Hao Zeng	Yuqian Zhang	Yushan Xue

15:20-15:50	**Tea Break**							
15:50-17:10	**Optimal Inference under Privacy, Communication, and Sampling Constraints**	**Advances in High-Dimensional Inference and Uncertainty Quantification**	**Regression and Time-series in High-dimensional Setting**	**Robust and Efficient Estimation for Modern Causal and Treatment Effect Models**	**Advances in Statistical Methods for Multiple Biomarkers and Data Sources**	**Recent Advances in Statistical Genetics**	**New Advances in Data Science**	**Multiple Testing Meets Modern Data and Machine Learning**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)	Shangwu Hall (商务厅)
	Chair: Lasse Vuursteen	Chair:Yin Xia	Chair:Yazhen Wang	Chair:Jia Gu	Chair: Xinzhou Guo	Chair:Shiyang Ma	Chair: Keren Li	Chair:Zinan Zhao
	Organizer: Lasse Vuursteen	Organizer: Yin Xia	Organizer: Yazhen Wang	Organizer: Matteo Bonvini	Organizer: Xinzhou Guo	Organizer: Shiyang Ma	Organizer: Zhigang Li& Lihui Zhao	Organizer: Zhimei Ren
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Mengchu Li	Jingyuan Liu	Shanshan Song	Wei Huang	Jialiang Li	Zilin Li	Tao Wang	Xiaoyang Wu
	Linjun Zhang	Zijian Guo	Qianqian Zhu	Mengchu Zheng	Wenlong Mou	Yaowu Liu	Zhonghua Liu	Yuan Yao
	Lasse Vuursteen	Juan Shen	Huiling Yuan	Lin Liu	Jiarui Zhang	Xianghong Hu	Jiang Gui	Bowen Gang
		Haojie Ren	Lan Li		Yao Zhang	Jingsi Ming	Keren Li	Zinan Zhao
18:00	**Banquet** Dongwan Hall（董湾厅)							

May 17th, Sunday							
09:00-09:50	Keynote Speech Richard J. Samworth (University of Cambridge)		Title: Outrigger local polynomial regression		Chair: Wenguang Sun		Dongwan Hall (董湾厅)
09:50-10:20	**Tea Break**						
10:20-11:40	**Model Averaging and Related Studies (MARS)**	**Statistical Learning on Complex Data Analysis**	**Statistical Learning Theory and Methods: From High-Dimensional Inference to Multi-Modal AI Applications**	**Recent Advances in Statistical Theory for Deep Learning**	**Frontiers in Interactive Generative World Models**	**Statistical Analysis of Network**	**Modern Methods in Statistical Learning and Data Analysis**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)
	Chair: Xinyu Zhang	Chair: Ting Li	Chair: Ling Zhou	Chair: Juntong Chen	Chair: Juyong Zhang	Chair: Yazhen Wang	Chair: Xintao Xia
	Organizer: Xinyu Zhang	Organizer: Ting Li	Organizer: Huazhen Lin	Organizer: Juntong Chen	Organizer: Juyong Zhang& Jianfei Cai	Organizer: Yazhen Wang	Organizer: Xintao Xia
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Xiangyu Cui	Zhenhua Lin	Wei Liu	Yuling Jiao	Lu Sheng	Yan Zhang	Ruijian Han
	Fang Fang	Wenliang Pan	Pengkun Yang	Huiming Zhang	Yudong Guo	Yongqin Qiu	Yu Gu
	Dandan Jiang	Long Feng	Lican Kang	Guohao Shen	Zhiwen Shao	Yi Ding	Chendi Wang
	Dalei Yu	Ting Li	Ling Zhou	Juntong Chen	Keyu Chen	Yuyang Liu	Yuxin Tao
11:40-14:00	**Lunch** Buffet Restaurant（Truffle）松露西餐厅						
14:00-15:20	**Recent Advances in the Statistical Foundations of Data Sciences**	**Statistical Inference and Sequential Decision Making**	**Deep Learning Methods in Survival Analysis**	**Recent Advances in Survival Analysis**	**Advances in Intelligent Agents and Efficient Multimodal Models**	**Preference-based centrality and ranking in general metrics space**	**Robust Treatment Effect Learning and Fair Decision Making**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)
	Chair: Anderson Ye Zhang	Chair: Kaijie Xue	Chair: Zhangsheng Yu	Chair: Zheng Chen	Chair: Bohan Zhuang	Chair: Tianxi Cai	Chair: Zheng Zhang
	Organizer: Rong Ma	Organizer: Anru Zhang	Organizer: Zhangsheng Yu	Organizer: Zheng Chen	Organizer: Bohan Zhuang& Jianfei Cai	Organizer: Tianxi Cai	Organizer: Wang Miao
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Anderson Ye Zhang	Yiyuan She	Xingqiu Zhao	Yongfu Yu	Chi Zhang	Doudou Zhou	Changcheng Li
	Zhengyu Huang	Bangyao Zhao	Wen Yu	Fangyao Chen	Wangbo Zhao	Mengyan Li	Molei Liu
	Weichen Wang	Ran Chen	Qixian Zhong	Huijuan Ma	Xu Yang	Yuming Zhang	Zeyu Bian
	Feiyu Jiang	Kaijie Xue	Zhangsheng Yu	Chengfeng Zhang	Bohan Zhuang	Alexander Giessing	Ping Zhang

15:20-15:50	**Tea Break**						
15:50-17:10	**Statistical Learning: Theory and Practice**	**Machine Learning and Statistical Inference for Complex Data**	**Next Generation Visual AIGC: Controllable Generation, Ultra-HD, and Intelligent Design**	**Advances in Statistical Learning and High-Dimensional Statistical Inference**	**Recent Advances in Causal Inference and Latent Variable Modeling**	**Advances in Inference for Modern Data Settings**	**Methods in Spatial Transcriptomic Data Analysis**
	Dongwan Hall A (董湾厅 A)	Dongwan Hall B (董湾厅 B)	Meishu Hall (梅墅厅)	Meizhu Hall (梅竹厅)	Jiaolu Hall A (茭芦厅 A)	Jiaolu Hall B (茭芦厅 B)	Yanshui Hall (烟水厅)
	Chair:Jing Zeng	Chair:Hanzhong Liu	Chair:Li Niu	Chair:Yeqing Zhou	Chair:Lin Liu	Chair:Tianxi Cai	Chair:Xin Yuan
	Organizer: Zhou Yu	Organizer: Hanzhong Liu	Organizer: Li Niu	Organizer: Yeqing Zhou	Organizer: Lin Liu	Organizer: Tianxi Cai	Organizer: Zhangsheng Yu
	Speakers:	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**	**Speakers:**
	Yuheng Ma	Jianqiao Wang	Xiaodong Cun	Jun Shu	Xin Zhang	Stéphane Guerrier	Chunman Zuo
	Jing Zeng	Xin Cong	Qian Yu	Danning Li	Zheng Zhang	Yanyuan Ma	Hongyi Xin
	Rui Qiu	Huaqing Jin	Ying Tai	Baihua He	Ruiyi Yang	Mucyo Karemera	Xin Yuan
	Ziwen Gao	Hanzhong Liu	Chunxia Xiao	Zhengtian Zhu	Yujia Gu	Luca Insolia	Ruitian Gao

Scientific Program

2026-05-16 Sat

Dongwan Hall

09:00-09:15 | Opening Ceremony
Chair: Zijian Guo

09:15-10:05 | Keynote Speech-Jun Liu
Chair: Tony Cai

09:15-10:05 Conditional sampling via diffusion flow and SMC
Jun Liu Tsinghua University

Dongwan Hall A

10:35-11:55 | Machine Learning for Policy Optimization and Causal Inference in Complex Systems
Organizer: Zijian Guo
Chair: Wei Luo

10:35-10:55 Adaptive Multi-task Learning for Multi-sector Portfolio Optimization
Qingliang Fan The Chinese University of Hong Kong

10:55-11:15 Optimizing Dynamic Treatment Regimes under Spatial Interference: Evidence from COVID-19 School Closures
Yunan Wu Yau Mathematical Sciences Center, Tsinghua University

11:15-11:35 Digital behaviors signal consumer well-being: A factor-augmented regularized prediction model approach
Xiangnan Feng Fudan University

11:35-11:55 MR2G: A novel framework for causal network inference using GWAS summary
Haoran Xue City University of Hong Kong

14:00-15:20 | Recent Advances in Statistics and AI
Organizer: Linjun Zhang
Chair: Linjun Zhang

14:00-14:20 A Statistical Framework for Alignment with Biased AI Feedback
Zhanrui Cai The University of Hong Kong

14:20-14:40 Labels or Preferences? Budget-Constrained Learning with Human Judgments over AI-Generated Outputs
Ruijia Wu Shanghai Jiao Tong University

14:40-15:00 Efficient machine unlearning with minimax optimality
Sai Li Tsinghua University

15:00-15:20 Anti-Concentration Inequalities for the Difference of Maxima of Gaussian Random Vectors
Shuting Shen National University of Singapore

15:50-17:10 | Optimal Inference under Privacy, Communication, and Sampling Constraints
Organizer: Lasse Vuursteen
Chair: Lasse Vuursteen

15:50-16:15 Federated Transfer Learning with Differential Privacy
Mengchu Li University of Birmingham

16:15-16:40 Evaluating LLMs When They Do Not Know the Answer: Statistical Evaluation of Mathematical Reasoning via Comparative Signals
Linjun Zhang Rutgers University

16:40-17:05 Optimality Theory for Adaptation under Differential Privacy
Lasse Vuursteen Duke University

Dongwan Hall B

10:35-11:55 | Advances in Causal Inference for Observation Data
Organizer: Zijian Guo
Chair: Zijian Guo

10:35-10:55 Confounder-robust causal discovery and inference in Perturb-seq using proxy and instrumental variables
Hongzhe Li University of Pennsylvania

10:55-11:15 Robust Bias Calibration for Causal Inference of Observational Data
Yumou Qiu Peking University

11:15-11:35 Correcting Nonignorable Nonresponse Bias in Turnout Estimation Using Callback Data
Wang Miao Peking University

11:35-11:55 Identification and Robust Inference for Multiple Treatments with Possibly Invalid Instruments
Ziwei Mei University of Macau

14:00-15:20 | Recent Developments in Hypothesis Testing
Organizer: Zijian Guo
Chair: Zijian Guo

14:00-14:20 Active Hypothesis Testing under Computational Budgets
Yin Xia Fudan University

14:20-14:40 Power Enhancement for Test of Multivariate Independence
Yeqing Zhou Tongji University

14:40-15:00 Model-X Knockoff Framework for Genome-Wide Survival Association Analysis
Shiyang Ma Shanghai Jiao Tong University School of Medicine

15:00-15:20 Is the F-test doubly robust?
Lucy Xia The Hong Kong University of Science and Technology

15:50-17:10 | Advances in High-Dimensional Inference and Uncertainty Quantification
Organizer: Yin Xia
Chair: Yin Xia

15:50-16:10 LLM-Powered Deep Panel Modeling
Jingyuan Liu Xiamen University

16:10-16:30 Statistical Inference for Conditional Group Distributionally Robust Optimization with Cross-Entropy Loss
Zijian Guo Zhejiang University

16:30-16:50 Variational Bayes for high-dimensional structured mixture model
Juan Shen Fudan University

16:50-17:10 Shape-Adaptive Conformal Prediction with Conditional Validity via Minimax Optimization
Haojie Ren Shanghai Jiao Tong University

Meizhu Hall

10:35-11:55 | Advances in Machine Learning Theory
Organizer: Jun Fan
Chair: Jun Fan

10:35-10:55 Towards Understanding How Transformers Perform In-Context Logistic Regression
Yuan Cao The University of Hong Kong

10:55-11:15 Non-vacuous Generalization Bounds for Overparameterized Shallow Neural Networks
Yunwen Lei *The University of Hong Kong*

11:15-11:35 Transformers Meet In-Context Learning: A Universal Approximation Theory
Gen Li *The Chinese University of Hong Kong*

11:35-11:55 Spectral Gradient Descent Mitigates Anisotropy-Driven Misalignment: A Case Study in Phase Retrieval
Wei Huang *RIKEN AIP and The Institute of Statistical Mathematics*

14:00-15:20 | Prediction and Machine learning in biomedical studies
Organizer: Zhigang Li
Chair: Jiang Gui

14:00-14:20 Machine learning methods to predict amyloid positivity using domain scores from cognitive tests
Guogen Shan *University of Florida*

14:20-14:40 Statistical methods for cell-cell interaction studies on spatially resolved transcriptomics
Xiaoyu Song *Duke-NUS Medical School*

14:40-15:00 Forecasting Respiratory Infectious Diseases with Adaptive and Physics-Informed Machine Learning
Kam Lun Tsang *The University of Hong Kong*

15:00-15:20 Joint modeling in presence of informative censoring on the retrospective time scale with application to palliative care research
Quran Wu *Jiangsu Hengrui Pharmaceuticals Co., Ltd.*

15:50-17:10 | Robust and Efficient Estimation for Modern Causal and Treatment Effect Models
Organizer: Matteo Bonvini
Chair: Jia Gu

15:50-16:15 Rate-Multiply Robust Estimation of General Treatment Models via Balanced Weighting
Wei Huang *University of Melbourne*

16:15-16:40 Perturbed Double Machine Learning: Nonstandard Inference Beyond the Parametric Length
Mengchu Zheng *Rutgers University*

16:40-17:05 Towards an efficiency theory on parameter manifolds
Lin Liu *Shanghai Jiao Tong University*

Meishu Hall

10:35-11:55 | Statistical and Machine Learning Methods for Analyzing Complex Real-world Data
Organizer: Rong Ma
Chair: Yiqun Chen

10:35-10:55 TBD
Yiqun Chen *Johns Hopkins University*

10:55-11:15 Latent Space Modeling for Human Disease Network with Temporal Variations: Analysis of Medicare Data
Hao Mei *Renmin University of China*

11:15-11:35 Causal effect heterogeneity estimation using summary statistics
Xingjie Shi *East China Normal University*

11:35-11:55 Statistical Inference for Random Objects
Hang Zhou *UNC Chape Hill*

14:00-15:20 | Recent Advances in Estimation and Inference for Structured and Heterogeneous Data
Organizer: Zhongyuan Lyu
Chair: Zhongyuan Lyu

14:00-14:25 High-dimensional Inference for Extreme Value Indices
Liujun Chen University of Science and Technology of China

14:25-14:50 Asymptotic Theory and Penalized Estimation for Signal-plus-Noise Matrices with Heteroskedastic Noise and Weak Signals
Zhixiang Zhang University of Macau

14:50-15:15 Probabilistic PCA on tensors
Yaoming Zhen The Chinese University of Hong Kong，Shenzhen

15:50-17:10 | Regression and Time-series in High-dimensional Setting
Organizer: Yazhen Wang
Chair: Yazhen Wang

15:50-16:10 Wasserstein Generative Regression
Shanshan Song Tongji University

16:10-16:30 A robust and scalable framework for high-dimensional volatility estimation
Qianqian Zhu Shanghai University of Finance and Economics

16:30-16:50 High-dimensional autoregressive time series modeling for symmetric matrices
Huiling Yuan East China Normal University

16:50-17:10 High-dimensional Autoregressive Modeling for Time Series Data with Hierarchical Structures
Lan Li The University of Hong Kong

Yanshui Hall

10:35-11:55 | From Statistical Foundations to AI Impacts: Risk, Robustness, and Representation
Organizer: Xin Tong
Chair: John Park

10:35-10:55 Quantifying Cross-Domain Knowledge Distillation in the Presence of Domain shift
Xiangchao Li University of Science and Technology of China

10:55-11:15 Geometric Fluctuations of Principal Subspaces for High-Dimensional Covariance Matrices
Yanlin Hu University of Science and Technology of China

11:15-11:35 Stance Drift: How AI-Mediated Communication Distorts Our Message
Lingchong Liu The Hong Kong University of Science and Technology

11:35-11:55 CalCS: Calibrated Cost-Sensitive Classification under Strict Error Constraints
John Park The University of Hong Kong

14:00-15:20 | Advanced Data Science Techniques: Differential Privacy, Dynamic Treatment Regimes
Organizer: Jia Gu
Chair: Jia Gu

14:00-14:20 Federated Learning of Quantile Inference under Local Differential Privacy
Shuyuan Wu Shanghai University of Finance and Economics

14:20-14:40 Minimax and Adaptive Covariance Matrix Estimation under Differential Privacy
Yicheng Li Tsinghua University

14:40-15:00 Communication-efficient Distributed Statistical Analysis under Differential Privacy
Haobo Qi Beijing Normal University

15:00-15:20 Balancing utility and cost in dynamic treatment regimes
Yuqian Zhang *Renmin University of China*

15:50-17:10 | New Advances in Data Science
Organizer: Zhigang Li & Lihui Zhao
Chair: Keren Li

15:50-16:10 Deciphering microbial community dynamics using cross-sectional data-informed NeuralODE
Tao Wang *Shanghai Jiao Tong University*

16:10-16:30 Constructive Instrumental Variable Identification and Inference with Many Weak Interaction Moments
Zhonghua Liu *Columbia University*

16:30-16:50 Words matter: Multimodal Suicide Risk Prediction from Veterans Health Administration Clinical Notes
Jiang Gui *Dartmouth College*

16:50-17:10 Distributed Learning with Heterogeneity and Asynchrony: Representative Learning
Keren Li *University of Alabama at Birmingham*

Jiaolu Hall A

10:35-11:55 | Statistical Theory and Random Matrices with Applications in Finance
Organizer: Zijian Guo
Chair: Xintao Xia

10:35-11:00 Cross-Sectional Learning and Inference for the Stochastic Discount Factor
Xinghua Zheng *HKUST*

11:00-11:25 Adaptive partial monitoring in non-stationary environments
Henry Reeve *Nanjing University*

11:25-11:50 Extremal eigenvectors of sparse random matrices
Yukun He *Fudan University*

14:00-15:20 | Private and Federated High-Dimensional Statistical Inference
Organizer: Dong Xia
Chair: Dong Xia

14:00-14:25 Differentially private sliced inverse regression in the federated paradigm
Xin Chen *Southern University of Science and Technology*

14:25-14:50 Adapting to noise tails in private linear regression
Lin Yang *Southwestern University of Finance and Economics*

14:50-15:15 Federated PCA: Differential Privacy, Algorithms, and Optimality under the Spiked Model
Dong Xia *Hong Kong University of Science and Technology*

15:50-17:10 | Advances in Statistical Methods for Multiple Biomarkers and Data Sources
Organizer: Xinzhou Guo
Chair: Xinzhou Guo

15:50-16:10 Weighted Youden Index Maximization
Jialing Li *National University of Singapore*

16:10-16:30 Provable imitation learning for nuclear fusion control
Wenlong Mou *University of Toronto*

16:30-16:50 Stein-Encoder: A White-Box Supervised Encoder via Stein Identities in Multi-Modal Studies
Jiarui Zhang *South China University of Technology*

16:50-17:10 Multi-Fidelity Quantile Regression
Yao Zhang *National University of Singapore*

Jiaolu Hall B

10:35-11:55 | Deep Generative Models and Their Interplay with Data-driven Decision-making
Organizer: Wenlong Mou
Chair: Wenlong Mou

10:35-10:55 Boosting as Sequential Aggregation of Generators for Structured Prediction
Jian Qian *University of Hong Kong*

10:55-11:15 New Frontiers on Statistical Learning with Prediction-Induced Distribution Shift
Zhun Deng *UNC at Chapel Hill*

11:15-11:35 A Queueing-Theoretic Framework for LLM Inference with KV Cache Memory Constraints
Nian Si *HKUST*

11:35-11:55 TBD
Yuchen Zhou *University of Illinois Urbana-Champaign*

14:00-15:20 | Transfer Learning and Robust Modeling for Structured Data
Organizer: Zijian Guo & Xintao Xia
Chair: Xintao Xia

14:00-14:20 Site Percolation Network Models for Event-Driven Systems
Yingying Li *HKUST*

14:20-14:40 Cross-Semanic Transfer Learning for High-dimensional Linear Regression
Xuejun Jiang *South University of Science and Technology of China*

14:40-15:00 Multi-Source Domain Adaptation via Alignment-Guided Distributionally Robust Learning
Zirui Wang *Tsinghua University*

15:00-15:20 Transfer Learning for Spatial Autoregressive Models with Application to U.S. Presidential Election Prediction
Hao Zeng *Southern University of Science and Technology*

15:50-17:10 | Recent Advances in Statistical Genetics
Organizer: Shiyang Ma
Chair: Shiyang Ma

15:50-16:10 All-in-One Toolkit for Biobank-Scale Whole-Genome Sequencing Data Management and Analysis
Zilin Li *Northeast Normal University*

16:10-16:30 A Powerful Transformation of Quantitative Responses for Biobank-Scale Association Studies
Yaowu Liu *Southwestern University of Finance and Economics*

16:30-16:50 A Unified Mendelian Randomization Framework for Identifying Causal Risk Factors: Accounting for Measured Covariates and Unmeasured Confounders
Xianghong Hu *Shenzhen Universit*

16:50-17:10 Bridging unpaired single-cell multimodal data for integrative analyses with SuperMap
Jingsi Ming *East China Normal University*

Shangwu Hall

10:35-11:55 | Statistical Learning for Complex Systems
Organizer: Shang Wu
Chair: Shang Wu

10:35-10:55 Computational and Statistical Asymptotic Analysis of the JKO Scheme with Unknown Parameters
Shang Wu Fudan University

10:55-11:15 Structural Change Detection in Dynamic Systems
Wei Zhang Fudan University

11:15-11:35 A Distributional Measure of Heterogeneous Variable Importance
Guorong Dai Fudan University

11:35-11:55 Opportunities and Challenges of Sharpness-Aware Training for Overparameterized Neural Networks
Chengli Tan Northwestern Polytechnical University

14:00-15:20 | Recent Advances in High-dimensional and Financial Statistics
Organizer: Yumou Qiu
Chair: Yumou Qiu

14:00-14:20 Asymptotics of higher criticism via Gaussian approximation
Jingkun Qiu Peking University

14:20-14:40 High-dimensional Clustering and Signal Recovery under Block Signals
Wu Su Peking University

14:40-15:00 Semiparametric Sieve Estimation for Survival Data with Two-layer Censoring
Yudong Wang University of Pennsylvania

15:00-15:20 Granular Data: Laying a Solid Foundation for High-Quality Datasets, Empowering AI Applications
Yushan Xue Central University of Finance and Economics

15:50-17:10 | Multiple Testing Meets Modern Data and Machine Learning
Organizer: Zhimei Ren
Chair: Zinan Zhao

15:50-16:10 Unified Conformalized Multiple Testing with Full Data Efficiency
Xiaoyang Wu Nankai University

16:10-16:30 Gold after Randomization: Model-X Split Knockoffs for FDR Control in Transformation Selection
Yuan Yao The Hong Kong University of Science and Technology

16:30-16:50 Multiple Testing Meets Data Visualization: A Modern Perspective on Boxplots and Bagplots
Bowen Gang Fudan University

16:50-17:10 A New Approach to Conformalized Model Selection for Out-of Distribution Testing
Zinan Zhao Zhejiang University

2026-05-17 Sun

Dongwan Hall

09:00-09:50 | Keynote Speech-Richard J. Samworth
Chair: Wenguang Sun

09:00-09:50 Outrigger local polynomial regression
Richard J. Samworth University of Cambridge

Dongwan Hall A

10:20-11:40 | Model Averaging and Related Studies (MARS)
Organizer: Xinyu Zhang
Chair: Xinyu Zhang

10:20-10:40 Policy Averaging for Stochastic Decision Problems: Theory and an Application to the Newsvendor Problem
Xiangyu Cui Shanghai University of Finance and Economics

10:40-11:00 Distributed propensity model averaging for large-scale data with nonignorable nonresponse
Fang Fang East China Normal University

11:00-11:20 Double Descent and Emergence in Multi-model Prediction
Dandan Jiang Xi'an Jiaotong University

11:20-11:40 Semi-supervised learning using copula-based regression and model averaging
Dalei Yu Xi'an Jiaotong University

14:00-15:20 | Recent Advances in the Statistical Foundations of Data Sciences
Organizer: Rong Ma
Chair: Anderson Ye Zhang

14:00-14:20 Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery
Anderson Ye Zhang University of Pennsylvania

14:20-14:40 Stable Gaussian Mixture Black-Box Variational Inference
Zhengyu Huang Peking University

14:40-15:00 Estimation of Out-of-Sample Sharpe Ratio for High Dimensional Portfolio Optimization
Weichen Wang Hong Kong University

15:00-15:20 Transfer learning in nonparametric online learning problems
Feiyu Jiang Fudan University

15:50-17:10 | Statistical Learning: Theory and Practice
Organizer: Zhou Yu
Chair: Jing Zeng

15:50-16:10 Locally Private Estimation with Public Features
Yuheng Ma Renmin University of China

16:10-16:30 Second-Order Sparse Sufficient Dimension Reduction with Applications to Quadratic Discriminant Analysis
Jing Zeng University of Science and Technology of China

16:30-16:50 Metric conformal prediction based on the expected local radius
Rui Qiu Peking University

16:50-17:10 Combining pre-trained large models via localized model averagin
Ziwen Gao Tsinghua University

Dongwan Hall B

10:20-11:40 | Statistical Learning on Complex Data Analysis
Organizer: Ting Li
Chair: Ting Li

10:20-10:40 Neural Wasserstein Two-Sample Tests
Zhenhua Lin National University of Singapore

10:40-11:00 Ball Impurity: Measuring Heterogeneity in General Metric Spaces
Wenliang Pan Academy of Mathematics and Systems Science, Chinese Academy of Sciences

11:00-11:20 Semi-Supervised Generative Learning via Latent Space Distribution Matching
Long Feng University of Hong Kong

11:20-11:40 Factor-augmented clustering tree for time series
Ting Li Southern University of Science and Technology

14:00-15:20 | Statistical Inference and Sequential Decision Making
Organizer: Anru Zhang
Chair: Kaijie Xue

14:00-14:20 Slacked Empirical Likelihoods for Post-Criterion Inference
Yiyuan She Westlake University

14:20-14:40 Bayesian reinforcement learning framework for optimizing the BCI-utility of P300 Brain-Computer Interfaces
Bangyao Zhao University of Michigan

14:40-15:00 Online Sequential Decision-Making with Reinforcement Learning: From Robotics to Human-centered tasks
Ran Chen Washington University in St. Louis

15:00-15:20 Inference on Large-scale Partially Functional Linear Model with Heterogeneous Errors
Kaijie Xue Shanghai University of International Business and Economics

15:50-17:10 | Machine Learning and Statistical Inference for Complex Data
Organizer: Hanzhong Liu
Chair: Hanzhong Liu

15:50-16:10 Heritability Estimation via Genetic Similarity Representation: Theory and Biobank-Scale Computation
Jianqiao Wang Tsinghua University

16:10-16:30 Generalizing Experience for Language Agents with Hierarchical MetaFlows
Xin Cong Tsinghua University

16:30-16:50 Conformal Inference for Minority Subgroups via Cross-Group Borrowing
Huaqing Jin Tsinghua University

16:50-17:10 Design-based inference for edge-level outcomes in directed networks
Hanzhong Liu Tsinghua University

Meizhu Hall

10:20-11:40 | Recent Advances in Statistical Theory for Deep Learning
Organizer: Juntong Chen
Chair: Juntong Chen

10:20-10:40 Inference-Time Alignment for Diffusion Models via Variationally Stable Doob's Matching
Yuling Jiao Wuhan University

10:40-11:00 Heavy-tailed Information-Theoretic Generalization Bounds with Applications to LLM Safety Alignment
Huiming Zhang Beihang University

11:00-11:20 Symmetry in Deep Neural Networks and Implications to Learning
Guohao Shen The Hong Kong Polytechnic University

11:20-11:40 A novel statistical approach to analyze image classification
Juntong Chen Xiamen University

14:00-15:20 | Recent Advances in Survival Analysis
Organizer: Zheng Chen
Chair: Zheng Chen

14:00-14:20 TBD
Yongfu Yu Fudan University

14:20-14:40 Doubly Robust Estimators for Heterogeneous Treatment Effects in Heteroskedastic Survival Data and Application
Fangyao Chen Xi'an Jiaotong University

14:40-15:00 Distributed Censored Quantile Regression: Convolution Smoothing and Communication Efficiency
Huijuan Ma East China Normal University

15:00-15:20 Dynamic-Centime：一种利用纵向数据预测剩余生存时间的动态预测模型
Chengfeng Zhang Southern Medical University

15:50-17:10 | Advances in Statistical Learning and High-Dimensional Statistical Inference
Organizer: Yeqing Zhou
Chair: Yeqing Zhou

15:50-16:10 模拟学习方法论：理论、算法及应用
Jun Shu Xi'an Jiaotong University

16:10-16:30 Strongly consistent community detection in popularity adjusted block models
Danning Li Northeast Normal University

16:30-16:50 Optimal Mixture-of-Experts Model Averaging for Conditional Generative Learning
Baihua He The University of Science and Technology of China

16:50-17:10 Controlling the False Discovery Rate in High-Dimensional Linear Models Using Model-X Knockoffs and p-values
Zhengtian Zhu Tongji University

Meishu Hall

10:20-11:40 | Statistical Learning Theory and Methods: From High-Dimensional Inference to Multi-Modal AI Applications
Organizer: Huazhen Lin
Chair: Ling Zhou

10:20-10:40 SMODER: Spatial Multi-Omics Deconvolution Anchored by RNA
Wei Liu Sichuan University

10:40-11:00 Nonparametric Inference on Unlabeled Histograms with Application to Generative Uncertainty Evaluation
Pengkun Yang Tsinghua University

11:00-11:20 Provable RLHF: A Consistent Framework for Offline Reward Learning and Value Function Learning
Lican Kang Wuhan University

11:20-11:40 Semi-supervised learning in high-dimensional linear regression
Ling Zhou Southwestern University of Finance and Economics

14:00-15:20 | Deep Learning Methods in Survival Analysis
Organizer: Zhangsheng Yu
Chair: Zhangsheng Yu

14:00-14:20 Identification and Inference for Structural Accelerated Failure Time Models via Instrument Interactions
Xingqiu Zhao The Hong Kong Polytechnic University

14:20-14:40 Neural frailty machines for survival analysis
Wen Yu Fudan University

14:40-15:00 Variable Significance Testing for the Deep Cox Model
Qixian Zhong Xiamen University

15:00-15:20 Deep partially linear transformation model for right-censored survival data
Zhangsheng Yu Shanghai Jiao Tong University

15:50-17:10 | Next Generation Visual AIGC: Controllable Generation, Ultra-HD, and Intelligent Design
Organizer: Li Niu
Chair: Li Niu

15:50-16:10 Towards Intelligent Story Visualization and Editing
Xiaodong Cun Great Bay University

16:10-16:30 Intelligent Design Generation: From SVG to Parametric CAD
Qian Yu Beihang University

16:30-16:50 Research on Data and Methods for Ultra-High-Definition Visual Generation
Ying Tai Nanjing University

16:50-17:10 Multi-View 3D Reconstruction with Radiance Fields
Chunxia Xiao Wuhan University

Yanshui Hall

10:20-11:40 | Modern Methods in Statistical Learning and Data Analysis
Organizer: Xintao Xia
Chair: Xintao Xia

10:20-10:40 Statistical ranking with dynamic covariates
Ruijian Han The Hong Kong Polytechnic University

10:40-11:00 Prediction-Oriented Transfer Learning for Survival Analysis
Yu Gu University of Hong Kong

11:00-11:20 What Separates Useful from Useless Synthetic Data? Verifying Synthetic Data through Representations
Chendi Wang Xiamen University

11:20-11:40 Homogeneity Pursuit in Ranking Inference Based on Pairwise Comparison
Yuxin Tao Southern University of Science and Technology

14:00-15:20 | Robust Treatment Effect Learning and Fair Decision Making
Organizer: Wang Miao
Chair: Zheng Zhang

14:00-14:20 Deep Learning Assisted Variable Selection with False Discovery Rate Control
Changcheng Li Dalian University of Technology

14:20-14:40 Maximin Learning of Individualized Treatment Effect on Multi-Domain Outcomes
Molei Liu Peking University

14:40-15:00 Double Fairness Policy Learning: Integrating Action Fairness and Outcome Fairness in Decision-making
Zeyu Bian FSU

15:00-15:20 Identifying the Desert Decision Rule to Assess and Achieve Fairness
Ping Zhang Peking University

15:50-17:10 | Methods in Spatial Transcriptomic Data Analysis
Organizer: Zhangsheng Yu
Chair: Xin Yuan

15:50-16:10 AI for dynamics network biology
Chunman Zuo Sun Yat-sen University

16:10-16:30 Local-and-Global Information-Preserving Statistical Manifold Learning for Single-Cell Transcriptomics
Hongyi Xin Shanghai Jiao Tong University

16:30-16:50 Deciphering Tissue Spatial Heterogeneity: Methods for the Identification of Spatially Variable Genes and the Characterization of Spatial Structures
Xin Yuan Shanghai Jiao Tong University

16:50-17:10 Meta-Encoder: Integrating Multiple Pathological Foundation Models to Enhance Predictive Accuracy for Key Cancer Biomarkers and Spatial Omics
Ruitian Gao Shanghai Jiao Tong University

Jiaolu Hall A

10:20-11:40 | Frontiers in Interactive Generative World Models
Organizer: Jianfei Cai & Juyong Zhang
Chair: Juyong Zhang

10:20-10:40 Towards 3D Content Generation and Understanding via Pretrained Generative Priors
Lu Sheng Beihang University

10:40-11:00 Realistic Human Interaction in Virtual Environments
Yudong Guo University of Science and Technology of China

11:00-11:20 Visual Affective-Cognitive Perception and Modeling
Zhiwen Shao China University of Mining and Technology

11:20-11:40 TBD
Keyu Chen Vivavia Inc.

14:00-15:20 | Advances in Intelligent Agents and Efficient Multimodal Models
Organizer: Bohan Zhuang &Jianfei Cai
Chair: Bohan Zhuang

14:00-14:20 Self-Evolving Agents: Boosting Efficiency and Capability via Architectural Progression
Chi Zhang Westlake University

14:20-14:40 Towards Efficient AI: Optimizing Deployment and Inference
Wangbo Zhao The Hong Kong University of Science and Technology

14:40-15:00 Analogy, Abstraction, and Reasoning in Multimodal Large Models
Xu Yang Southeast University

15:00-15:20 Towards Efficient Inference of Large Foundation Models
Bohan Zhuang Zhejiang University

15:50-17:10 | Recent Advances in Causal Inference and Latent Variable Modeling
Organizer: Lin Liu
Chair: Lin Liu

15:50-16:10 Bias reduction in g-computation for covariate adjustment in randomized clinical trials
Xin Zhang Pfizer Inc.

16:10-16:30 Higher-order debiased estimators of general treatment models
Zheng Zhang Renmin University of China

16:30-16:50 Model-free Estimation of Latent Structure via Multiscale Nonparametric Maximum Likelihood
Ruiyi Yang Shanghai Jiao Tong University

16:50-17:10 Incorporating external data for analyzing randomized clinical trials: A transfer learning approach
Yujia Gu Tsinghua University

Jiaolu Hall B

10:20-11:40 | Statistical Analysis of Networks
Organizer: Yazhen Wang
Chair: Yazhen Wang

10:20-10:40 Common-Individual Embedding for Dynamic Networks with Temporal Group Structure
Yan Zhang Shanghai University of International Business and Economics

10:40-11:00 Latent Space Model under Edge Contamination
Yongqin Qiu University of Science and Technology of China

11:00-11:20 Likelihood-Based Change Point Detection for Sparse SBM Networks
Yi Ding School of Statistics, University of International Business and Economics

11:20-11:40 Graphical Models for Mixed-type Data
Yuyang Liu Shanghai University of International Business and Economics

14:00-15:20 | Preference-based centrality and ranking in general metrics space
Organizer: Tianxi Cai
Chair: Tianxi Cai

14:00-14:20 Hierarchical Contrastive Learning for Multimodal Data with Partial Sharing
Doudou Zhou National University of Singapore

14:20-14:40 Semi-supervised Clustering Through Representation Learning of High-dimensional Count Data
Mengyan Li Bentley University

14:40-15:00 Multi-view Spherical Mixture Models for Aligning Partially Overlapping Feature Spaces with Latent Synonymy
Yuming Zhang Harvard University

15:00-15:20 Efficient inference on high-dimensional logistic regression under class imbalance
Alexander Giessing National University of Singapore

15:50-17:10 | Advances in Inference for Modern Data Settings
Organizer: Tianxi Cai
Chair: Tianxi Cai

15:50-16:10 Bias Correction for Semiparametric Regression Models
Stéphane Guerrier University of Geneva

16:10-16:30 Studies in Label Shift
Yanyuan Ma PSU

16:30-16:50 The implicit bootstrap: a percentile second-order correct interval estimation method
Mucyo Karemera University of Geneva

16:50-17:10 Partial optimality and applicability for multivariate equivalence testing
Luca Insolia University of Geneva

Keynote Speech

Jun Liu

Tsinghua University

Dr. Jun Liu is Xinghua Distinguished University Professor, Chair of the Department of Statistics and Data Science at Tsinghua University, and a member of the National Academy of Sciences of the USA. He received his BS degree in mathematics in 1985 from Peking University and Ph.D in statistics in 1991 from the University of Chicago. From 1991-2025, he held Assistant, Associate, and Full professorship at Harvard and Stanford Universities. Liu won the COPSS Presidents' Award in 2002, the Morningside Gold Medal in Applied Mathematics in 2010, and the Pao-Lu Hsu Award by ICSA in 2016. He was elected to Fellow of IMS, ASA, and ISCB in 2004, 2005, and 2022, respectively, and to the National Academy of Sciences of the USA in 2025. Liu has served as the co-editor for the flagship statistics journal JASA from 2011-2014, as associate editor for leading statistical journals, and as a committee chair or member for various grant review panels. Dr. Liu has co-authored over 300 research articles published in leading scientific journals, conferences and books, with a Google Scholar citation count of more than 97,000. Over the past four decades, he has mentored more than 40 PhD students and 30 postdoctoral fellows. Liu's research interests are: Bayesian methods and computation, statistical machine learning and AI, Monte Carlo methods, state-space models, bioinformatics and computational biology.

Time: May 16, 09:15-10:05
Venue: Dongwan Hall

Title:Conditional sampling via diffusion flow and SMC

Abstract:
Sequential Monte Carlo, aka particle filtering, refers to a class of Monte Carlo methods that accommodates dynamic structures and can be used as a learning mechanism. The scheme starts by creating the sampling distribution recursively and adjusting the obtained samples by sequentially adjusted weights so as to"learn"when new information is available. Recently, diffusion models have become a very popular tool for learning a high-dimensional data-distribution and generating from it. I will review a brief history of SMC and some developments in diffusion modeling. By combining the ODE-based flow method and SMC, we propose a training-free conditional sampling method for diffusion models. Because a naive application of importance sampling suffers from weight degeneracy in high-dimensional settings, ideas of resampling and rejection sampling are necessary. To encourage generated samples to diverge along distinct trajectories, we derive a stochastic flow with adjustable noise strength to replace the deterministic flow at the intermediate stage. Experimentally, our method significantly outperforms existing approaches on conditional sampling tasks for MNIST and CIFAR-10.

Richard J. Samworth

University of Cambridge

Richard Samworth obtained his PhD in Statistics from the University of Cambridge in 2004, and has remained in Cambridge since, becoming a full professor in 2013 and the Professor of Statistical Science in 2017. His main research interests are in nonparametric and high-dimensional statistics, as well as the statistical foundations of AI; he has developed methods and theory for shape-constrained inference, missing data, subgroup selection, deep learning, data perturbation techniques, changepoint estimation, variable selection and independence testing. Richard received the COPSS Presidents' Award in 2018, was elected as a Fellow of the Royal Society in 2021 and was awarded the David Cox Medal for Statistics in 2025. He served as co-editor of the Annals of Statistics (2019-2021) and is currently IMS President-Elect.

Time: May 17, 09:00-09:50
Venue: Dongwan Hall

Title: Outrigger local polynomial regression

Abstract:
Standard local polynomial estimators of a nonparametric regression function employ a weighted least squares loss function that is tailored to the setting of homoscedastic Gaussian errors. We introduce the outrigger local polynomial estimator, which is designed to achieve distributional adaptivity across different conditional error distributions. It modifies a standard local polynomial estimator by employing an estimate of the conditional score function of the errors and an 'outrigger' that draws on the data in a broader local window to stabilise the influence of the conditional score estimate. Subject to smoothness and moment conditions, and only requiring consistency of the conditional score estimate, we first establish that even under the least favourable settings for the outrigger estimator, the asymptotic ratio of the worst-case local risks of the two estimators is at most 1, with equality if and only if the conditional error distribution is Gaussian. Moreover, we prove that the outrigger estimator is minimax optimal over Hölder classes up to a multiplicative factor $A_{\beta,d}$, depending only on the smoothness $\beta \in (0,\infty)$ of the regression function and the dimension d of the covariates. When $\beta \in (0,1]$, we find that $A_{\beta,d} \leq 1.69$, with $\lim_{\beta \searrow 0} A_{\beta,d} =1$. A further attraction of our proposal is that we do not require structural assumptions such as independence of errors and covariates, or symmetry of the conditional error distribution. Numerical results on simulated and real data validate our theoretical findings; our methodology is implemented in the R package \texttt{outrigger}.

Abstracts

Adaptive Multi-task Learning for Multi-sector Portfolio Optimization

Qingliang Fan

The Chinese University of Hong Kong

Abstract： Accurately capturing the transfer of information across multiple sectors to enhance model estimation is both significant and challenging in multi-sector portfolio optimization involving a large number of assets in different sectors. Within the framework of factor modeling, we propose a novel data-adaptive multi-task learning methodology that quantifies and learns the relatedness among the principal temporal subspaces (spanned by factors) across multiple sectors. This approach improves the estimation of multiple factor models and thus enhances multi-sector portfolio optimization. A novel and easy-to-implement algorithm, termed projection-penalized principal component analysis, is developed to accomplish the multi-task learning procedure. We establish asymptotic properties for both the estimators from the multi-task factor model and the associated multi-task portfolio risk and Sharpe ratio estimators. Our simulation study, and empirical applications based on the data of the Russell 3000 index components with the Fama-French industrial sectors, demonstrate the advantages of the newly proposed multi-task learning methodology.

Optimizing Dynamic Treatment Regimes under Spatial Interference: Evidence from COVID-19 School Closures

Yunan Wu

Tsinghua University

Abstract： Effective dynamic decision-making plays a central role in policy formation, yet standard approaches often neglect spatial interdependencea fundamental feature of real-world settings. In practice, policies are implemented across interconnected states or regions, where outcomes in one unit may be influenced by interventions in others. To address this limitation, we develop a novel framework for estimating optimal dynamic treatment regimes under spatial interference. We propose methods to estimate optimal sequences of policy decisions over multiple time points while explicitly incorporating spatial spillover effects. For statistical inference, we construct confidence intervals for the parameters indexing the optimal regime using multiplier bootstrap techniques, and rigorously establish the theoretical validity of the proposed procedures. We demonstrate the practical utility of our framework through an application to COVID-19 school closure policies in the United States. Using state-level intervention and outcome data, we show that accounting for spatial dependence can lead to more effective policy decisions. These results highlight the importance of incorporating network structure into policy design and provide a principled approach for developing more precise and context-aware intervention strategies in interconnected populations.

Digital behaviors signal consumer well-being: A factor-augmented regularized prediction model approach

Xiangnan Feng

Fudan University

Abstract: We examine whether digital behaviors signal consumer well-being by linking individual-level mobile-phone app-usage data to surveys of the same users' happiness and stress. Our main study context is (both before and during) the COVID-19 pandemic outbreak in a major Chinese city. We use verifiable digital-trace data to document lifestyle shifts during the lockdown, as well as age and gender heterogeneities in behavioral change. We propose the use of an advanced factor-augmented regularized prediction model (FarmPredict) to exploit the mobile app usage patterns for predicting happiness and stress before and after the pandemic outbreak. Mobile app usage data contain useful signals: Both latent factors and idiosyncratic residuals extracted from app usage by a factor model contribute to predicting happiness and stress; the common factors meaningfully capture app usage patterns and digital lifestyles, while idiosyncratic residuals carry valuable individualized information. Moreover, a first-difference analysis identifies the app categories and latent factors that predicted changes in well-being during the pandemic. To assess the ecological validity of the model with a different population and stressor, we further validate the findings in a separate lab experiment using an undergraduate student sample whose subjective well-being fluctuates during the finals season. Across contexts, individuals' digital behaviors coupled with FarmPredict function as a population-level "hedonometer": They provide a continuous, passive indicator of well-being, clarify the roles of common lifestyle factors versus idiosyncratic residuals, and help identify psychologically vulnerable subpopulations.

MR2G: A novel framework for causal network inference using GWAS summary

Haoran Xue

City University of Hong Kong

Abstract: Inferring a causal network among multiple traits is essential for unraveling complex biological relationships and informing interventions. Mendelian randomization (MR) has emerged as a powerful tool for causal inference, utilizing genetic variants as instrumental variables (IVs) to estimate causal effects. However, when the directions of causal relationships among traits are unknown, reconstructing the underlying causal network becomes challenging. In particular, the presence of cycles or feedback loops, which are common in biological systems, poses additional challenges for causal network inference, and remains largely under-studied with standard MR approaches and existing IV-based network inference methods. To address these issues, we introduce MR2G, a new statistical framework that enables robust inference of causal networks, including those with cycles, directly from GWAS summary statistics. MR2G is built on a formally defined recursive causal graph model that rigorously links direct causal effects to MR estimands. It recovers a biologically interpretable causal network from pairwise MR effect estimates, while incorporating a network-informed IV screening strategy to reduce pleiotropic bias and improve robustness. Through realistic simulations, MR2G demonstrates superior accuracy and robustness in recovering complex causal structures, including those involving feedback loops. We apply MR2G to GWAS summary statistics for six complex diseases and nine cardiometabolic risk factors. MR2G not only recovers well-established causal pathways but also uncovers multiple feedback relationships, highlighting its utility in disentangling complex and biologically plausible causal networks from large-scale genetic data.

Confounder-robust causal discovery and inference in Perturb-seq using proxy and instrumental variables

Hongzhe Li

University of Pennsylvania

Abstract： Emerging single-cell technologies that integrate CRISPR-based genetic perturbations with single-cell RNA sequencing, such as Perturb-seq, have substantially advanced our understanding of gene regulation and causal influence of genes. While Perturb-seq data provide valuable causal insights into gene-gene interactions, statistical concerns remain regarding unobserved confounders that may bias inference. These latent factors may arise not only from intrinsic molecular features of regulatory elements encoded in Perturb-seq experiments, but also from unobserved genes arising from cost-constrained experimental designs. Although methods for analyzing large-scale Perturb-seq data are rapidly maturing, approaches that explicitly account for such unobserved confounders in learning the causal gene networks are still lacking. Here, we propose a novel method to recover causal gene networks from Perturb-seq experiments with robustness to arbitrarily omitted confounders. Our framework leverages proxy and instrumental variable strategies to exploit the rich information embedded in perturbations, enabling unbiased estimation of the underlying directed acyclic graph (DAG) of gene expressions. Simulation studies and analyses of CRISPR interference experiments of K562 cells demonstrate that our method outperforms baseline approaches that ignore unmeasured confounding, yielding more accurate and biologically relevant recovery of the true gene causal DAGs.

Robust Bias Calibration for Causal Inference of Observational Data

Yumou Qiu

Peking University

Abstract： Residual systematic bias from unmeasured confounders remains a central challenge of causal inference in observational studies. Negative control outcomes (NCOs) are widely used for bias calibration, but their validity cannot be guaranteed a priori and invalid controls can distort calibration. We introduce reference outcomes (ROs), requiring only equality of population mean potential outcomes across treatment arms, and develop a robust distributional calibration method allowing an unknown fraction of candidate ROs to be invalid. Using only summary-level inputs, specifically, effect estimates and their variance estimates, we propose a convolved mixture in which valid ROs share a systematic-bias distribution while invalid ROs follow its convolution with a nonzero-effect distribution. Intrinsic asymmetry yields nonparametric identifiability of the bias distribution without majority-validity assumptions. Within parametric families, we develop a maximum pseudo-likelihood estimator with an EM algorithm that produces (i) bias-adjusted confidence intervals for the primary outcome and (ii) posterior invalidity probabilities for candidate ROs, enabling transparent diagnostics. We establish consistency, asymptotic normality, and asymptotic coverage under a joint asymptotic regime. An application to TriNetX comparing GLP-1RAs versus SGLT2is for arrhythmic outcomes demonstrates robust calibration using either crude or expert-refined RO lists, reducing reliance on exact prior knowledge of RO validity and extensive expert screening.

Correcting Nonignorable Nonresponse Bias in Turnout Estimation Using Callback Data

Wang Miao

Peking University

Abstract: Overestimation of turnout has long been an issue in election surveys, with nonresponse bias or voter overrepresentation identified as major sources of bias. However, adjusting for nonignorable nonresponse bias is substantially challenging. Based on the ANES Non-Response Follow-Up study concerning the 2020 U.S. presidential election, we investigate the role of callback data, that is, records of contact attempts in the survey course, in adjusting for nonresponse bias in the estimation of turnout. We propose a stableness of resistance assumption to account for nonignorable missingness in the outcome, which states that the impact of the missing outcome on the response propensity is stable in the first two call attempts. Under this assumption and by integrating with covariate information from the census data, we establish identifiability and develop estimation methods for turnout. Our methods produce estimates very close to the official turnout and successfully capture the trend of declining willingness to vote as response reluctance increases. This work highlights the importance of adjusting for nonignorable nonresponse bias and demonstrates the potential of widely available callback data for political surveys.

Identification and Robust Inference for Multiple Treatments with Possibly Invalid Instruments

Ziwei Mei

University of Macau

Abstract: The instrumental variable (IV) method is widely used to infer causal effects in observational studies with unmeasured confounding. Causal identification becomes challenging when some IVs are possibly invalid. Unlike existing work that focuses on a single treatment, we develop identification conditions and robust inference procedures for multiple endogenous treatments with possibly invalid instruments. Identification and inference with multiple treatments are more challenging than those in single-treatment settings because point identification of multidimensional treatment effects requires multiple valid instruments. We introduce a necessary and sufficient condition, the generalized plurality rule, for identifying multidimensional treatment effects with possibly invalid IVs. We further exploit this condition to select valid instruments in a data-dependent way. However, local-to-zero violations of IV validity can induce under-coverage when the selected set includes locally invalid IVs. To address this, we propose establishing confidence intervals for multiple treatments using a sampling method that is robust to instrument-selection errors and retains nominal coverage. We demonstrate the usefulness of our inferential method through simulations and an application to Mendelian randomization.

TBD

Yiqun Chen

Johns Hopkins University

Abstract：TBD

Latent Space Modeling for Human Disease Network with Temporal Variations: Analysis of Medicare Data

Hao Mei[1],Ruiyue Wang[1],Rong Li[2],Sanguo Zhang[1],Shuangge Ma[1],Guangzhong Qiao[3],Hao Mei[4]

1. School of Mathematical Sciences, University of Chinese Academy of Science
2. Department of Biostatistics, Yale School of Public Health
3. Department of Orthopaedic, The First Hospital of Tsinghua University
4. Center for Applied Statistics, School of Statistics, Institute of Health Data Science, Renmin University of China

Abstract： Human disease network (HDN) analysis, which jointly considers a large number of diseases and focuses on their interconnections, is getting increas_x0002_ingly popular and can shed important insight not possessed by individual_x0002_disease-based analysis. Multiple network analysis techniques have been de_x0002_veloped for HDNs, although new developments are still strongly needed. In this article we adopt latent space modeling, which has proven powerful in other network analysis contexts and offers unique, insightful interpretations, but has been limitedly applied in HDN analysis. Different from some other types of network analysis and some other HDN analyses (such as gene-centric ones), in this article we pay unique attention to modeling temporal variations. For this purpose, a penalization approach is developed, which can identify time regions with constant network structures (that correspond to ignorable changes) as well as those with smooth variations. The statistical and com_x0002_putational properties are rigorously established. With Medicare data—one of the most powerful medical claims databases—we analyze the admission records of 133 million hospital inpatient treatments from January 2008 to December 2019. Sensible findings are made on disease interconnections and clustering structures. Additionally, the temporal variations, which have not been revealed in the literature, are found to be interpretable. The analysis can provide a new way for connecting and grouping diseases and assist in under_x0002_standing and planning medical resources.

Causal effect heterogeneity estimation using summary statistics

**Xingjie Shi[1],Minxi Bai[1],Jiacheng Miao[2],Stephen Dorn[2],Jonathan Haugstad[2],
Jin Liu[3],Qiongshi Lu[2],Xingjie Shi[1]**

1. KLATASDS-MOE, Academy of Statistics and Interdisciplinary Sciences, School of Sta-tistics, East China Normal University
2. Department of Biostatistics & Medical Informatics, University of Wisconsin–Madison
3. School of Data Science, The Chinese University of Hong Kong, Shenzhen

Abstract： Mendelian randomization (MR) has swiftly gained popularity as a tool for causal inference in genetic epidemiology. However, existing MR methods focus exclusively on estimating the average causal effect and cannot quantify its heterogeneity, posing a major methodological limitation and impeding context-dependent causal findings. Here, we introduce MEndelian Randomization for Linear INteraction (MERLIN), a unified Bayesian framework that jointly estimates the average and context-dependent causal effects using summary data from genome-wide association and interaction studies. Through extensive simulation analyses, we demonstrate the improved power, robustness, and broad utility of MERLIN versus existing methods. We show MERLIN was able to identify sex-specific causal effects of schizophrenia on brain imaging traits, a male-specific causal effect of testosterone on bipolar disorder, and age-dependent causal effects of metabolic biomarkers on coronary artery disease risk. These results illustrate the transformative potential of summary-data-based inference for causal heterogeneity. Together, MERLIN provides a powerful and practical framework for investigating causal effect heterogeneity using summary-level observational data and greatly enhances our capability to elucidate complex disease etiology.

Statistical Inference for Random Objects

Hang Zhou

UNC

Abstract： Random objects are complex random variables taking values in general metric spaces. Although such data are increasingly common in scientific research, current statistical methodology and theory remain limited. The primary challenge in analyzing such data lies in the absence of vector space operations, such as addition, subtraction, scalar multiplication, and inner products, which are fundamental tools in conventional statistical methodologies. This talk explores object data with distance profiles and their application to conformal prediction and independence testing.

Towards Understanding How Transformers Perform In-Context Logistic Regression

Yuan Cao

The University of Hong Kong

Abstract: Transformers have demonstrated remarkable in-context learning (ICL) capabilities. The strong ICL performance of transformers is commonly believed to arise from their ability to implicitly execute certain algorithms on the context, thereby enhancing prediction and generation. In this work, we investigate how transformers with softmax attention perform in-context learning on linear classification data. We first construct a class of multi-layer transformers that can perform in-context logistic regression, with each layer exactly performing one step of normalized gradient descent on an in-context loss. Then, we show that our constructed transformer can be obtained through (i) training a single self-attention layer supervised by one-step gradient descent, and (ii) recurrently applying the trained layer to obtain a looped model. Training convergence guarantees of the self-attention layer and out-of-distribution generalization guarantees of the looped model are provided. Our results advance the theoretical understanding of ICL mechanism by showcasing how softmax transformers can effectively act as in-context learners.

Non-vacuous Generalization Bounds for Overparameterized Shallow Neural Networks

Yunwen Lei

The University of Hong Kong

Abstract: Overparameterized neural networks often exhibit a benign overfitting phenomenon, achieving excellent generalization performance despite having more parameters than training samples. Traditional generalization analyses typically yield vacuous bounds due to overparameterization, which cannot explain the extraordinary generalization behavior of overparameterized neural networks in practice. In this talk, we establish non-vacuous generalization bounds by controlling the Rademacher complexity of overparameterized shallow neural networks (SNNs), supported by empirical studies involving highly overparameterized SNNs. Our complexity bounds are fully dependent on the distance from the initialization point and are expressed in terms of the path-norm of the networks. Experimental analyses show that our theoretical analysis implies non-vacuous generalization bounds even if the model is highly overparameterized.

Transformers Meet In-Context Learning: A Universal Approximation Theory

Gen Li

The Chinese University of Hong Kong

Abstract: Large language models are capable of in-context learning, the ability to perform new tasks at test time using a handful of input-output examples, without parameter updates. We develop a universal approximation theory to elucidate how transformers enable in-context learning. For a general class of functions (each representing a distinct task), we demonstrate how to construct a transformer that, without any further weight updates, can predict based on a few noisy in-context examples with vanishingly small risk. Unlike prior work that frames transformers as approximators of optimization algorithms (e.g., gradient descent) for statistical learning tasks, we integrate Barron's universal function approximation theory with the algorithm approximator viewpoint. Our approach yields approximation guarantees that are not constrained by the effectiveness of the optimization algorithms being mimicked, extending far beyond convex problems like linear regression. The key is to show that (i) any target function can be nearly linearly represented, with small l1-norm, over a set of universal features, and (ii) a transformer can be constructed to find the linear representation -- akin to solving Lasso -- at test time.
This is joint work with Yuchen Jiao, Yu Huang, Yuting Wei, and Yuxin Chen.

Spectral Gradient Descent Mitigates Anisotropy-Driven Misalignment: A Case Study in Phase Retrieval

Wei Huang

RIEKN AIP

Abstract: Spectral gradient methods, such as the Muon optimizer, modify gradient updates by preserving directional information while discarding scale, and have shown strong empirical performance in deep learning. We investigate the mechanisms underlying these gains through a dynamical analysis of a nonlinear phase retrieval model with anisotropic Gaussian inputs, equivalent to training a two-layer neural network with the quadratic activation and fixed second-layer weights. Focusing on a spiked covariance setting where the dominant variance direction is orthogonal to the signal, we show that gradient descent (GD) suffers from a variance-induced misalignment: during the early escaping stage, the high-variance but uninformative spike direction is multiplicatively amplified, degrading alignment with the true signal under strong anisotropy. In contrast, spectral gradient descent (SpecGD) removes this spike amplification effect, leading to stable alignment and accelerated noise contraction. Numerical experiments confirm the theory and show that these phenomena persist under broader anisotropic covariances.

Cross-Sectional Learning and Inference for the Stochastic Discount Factor

Zhanhui Chen,Yi Ding,Yingying Li,Xinghua Zheng,Xinghua Zheng

HKUST

Abstract： We develop a statistical learning framework for constructing the stochastic discount factor (SDF) portfolio. To address the dimensionality challenge, we extend the MAXSER method (Ao, Li and Zheng, 2019) to allow for N >> T; prove that it surely screens for useful characteristics; and establish asymptotic normality for the SDF loading estimates. Using 153 characteristics returns from cross-sectional regressions (Fama and French, 2020), our framework not only constructs an SDF with a high out-of sample Sharpe ratio that successfully prices the cross-section of expected returns, but also allows us to identify key characteristic themes and test the significance of their contributions.

Adaptive partial monitoring in non-stationary environments

Henry Reeve

Nanjing University

Abstract： We introduce a flexible framework for sequential decision making in a non-stationary stochastic environment. The regret of a policy contrasts performance with the expected reward of a dynamic oracle capable of selecting an optimal sequence of actions for the non-stationary stochastic environment. We introduce an algorithm which leverages e-processes to provably adapt to distributional changes in settings where the reward attained from a given action is not directly observed. We demonstrate that the optimal regret depends upon a fascinating interplay between the level of observability, the noise level, the complexity of the action space, and the degree of non-stationarity.

Extremal eigenvectors of sparse random matrices

Yukun He

Fudan University

Abstract: We consider a class of sparse random matrices, which includes the adjacency matrix of the Erd˝os-R ´enyi graph G(N, p). For N^{−1+o(1)} ⩽ p ⩽ 1/2, we show that the non-trivial edge eigenvectors are asymptotically jointly normal. The main ingredient of the proof is an algorithm that directly computes the joint eigenvector distributions, without comparisons with GOE. The method is applicable in general. As an illustration, we also use it to prove the normal fluctuation in quantum ergodicity at the edge for Wigner matrices. Another ingredient of the proof is the isotropic local law for sparse matrices, which at the same time improves several existing results.
This is joint work with Jiaoyang Huang and Chen Wang.

Boosting as Sequential Aggregation of Generators for Structured Prediction

Jian Qian[1],Shu Ge[2]

1. University of Hong Kong
2. Independent Researcher

Abstract： Despite the recent success of deep generative models for structured prediction, including approaches based on ensembling multiple models, a general theoretical understanding of how to effectively combine generators remains limited.

In this talk, we present a boosting framework for structured prediction, providing theoretical guarantees for aggregating generators into a stronger one. Our approach proceeds via sequential reweighting of training samples, where each round focuses on correcting the current model's errors under adaptively chosen data distributions.

A key contribution is the identification of an(alpha,beta)-stability condition, which characterizes when aggregation can successfully amplify weak guarantees. We show that this stability critically depends on the choice of divergence used to evaluate prediction quality. In particular, under suitable divergences, if generators exist with respect to a sequence of adaptively reweighted samples, then their aggregation yields a stronger generator with provably small total loss.

These results provide a unified perspective on boosting as a principled method for combining generators, highlighting the role of divergence geometry and adaptive data reweighting in structured prediction.

New Frontiers on Statistical Learning with Prediction-Induced Distribution Shift

Zhun Deng[1],Kaicheng Zhang[1],Lihua Lei[2],Zhun Deng[1]

1. UNC at Chapel Hill
2. Stanford University

Abstract： Prediction-induced distribution shift refers to the phenomenon where prediction-informed decisions alter the data-generating process underlying the outcomes they aim to predict. This type of distribution shift is pervasive in both machine learning and policy-making. For example, loan policies based on default risk predictions can alter consumer behavior in ways that in turn affect repayment outcomes. While a rich body of computer science literature has studied convergence and optimization under prediction-induced distribution shift, the reliability of decision-making in this setting remains severely underexplored.

In this talk, we will present two new studies that explore reliability and uncertainty in decision-making systems where predictions interact with human behavior: inference under prediction-induced distribution shift and human–AI collaboration under AI-induced framing effects.

First, we will introduce a framework for statistical inference under prediction-induced distribution shift. Our contributions are twofold: we establish a central limit theorem for estimation and inference in this setting, and extend it to prediction-powered inference, which leverages a small labeled dataset and large-scale predictions to obtain more precise estimates and sharper confidence regions for policy-relevant parameters.

Second, we will introduce a framework for studying human–AI collaboration under AI-induced framing effects. In many decision-making settings, AI systems do not merely provide information to human experts; they also shape how experts interpret evidence, express uncertainty, and exercise their own judgment. We study how different framings of AI outputs can affect human expertise, including when experts defer to, discount, or recalibrate their judgments in response to AI-generated recommendations. Through this lens, we examine how uncertainty quantification can support more reliable human–AI collaboration by helping experts appropriately interpret AI assistance rather than being inadvertently steered by it.

Together, these studies highlight new challenges and opportunities at the intersection of machine learning, AI, statistics, and point toward principled ways to handle uncertainty when predictions shape both real-world outcome distributions and human expert decisions.

A Queueing-Theoretic Framework for LLM Inference with KV Cache Memory Constraints

Chengyi Nie[2],Nian Si[1],Zijie Zhou[1]

1. HKUST
2. State University of New York at Stony Brook

Abstract： The rapid adoption of large language models (LLMs) has created significant challenges for efficient inference at scale. Unlike traditional workloads, LLM inference is constrained by both computation and the memory overhead of key–value (KV) caching, which accelerates decoding but quickly exhausts GPU memory. In this paper, we introduce the first queueing-theoretic framework that explicitly incorporates both computation and GPU memory constraints into the analysis of LLM inference. Based on this framework, we derive rigorous stability and instability conditions that determine whether an LLM inference service can sustain incoming demand without unbounded queue growth. This result offers a powerful tool for system deployment, potentially addressing the core challenge of GPU provisioning. By combining an estimated request arrival rate with our derived stable service rate, operators can calculate the necessary cluster size to avoid both costly over-purchasing and performance-violating under-provisioning. We further validate our theoretical predictions through extensive experiments in real GPU production environments. Our results show that the predicted stability conditions are highly accurate, with deviations typically within 10%.

TBD

Yuchen Zhou

University of Illinois Urbana-Champaign

Abstract：TBD

Quantifying Cross-Domain Knowledge Distillation in the Presence of Domain shift

Xiangchao Li

University of Science and Technology of China

Abstract: This paper presents a theoretical investigation into the generalization capabilities of cross-domain knowledge distillation. Utilizing a high-dimensional asymptotic analysis of a linear teacher–student model, we characterize the excess risk while accounting for both model and covariate shifts. Our results provide a formal guarantee for the efficacy of distillation: even when the source and target domains differ substantially, there still may exist a regime where the student model achieves superior generalization ability over the student-only baseline. Moreover, we identify a crossed double descent phenomenon: the excess risk can vary non-monotonically with the teacher's and student's dimension-to-sample-size ratios. These results provide rigorous insight into when and why distillation helps across domains.

Geometric Fluctuations of Principal Subspaces for High-Dimensional Covariance Matrices

Yanlin Hu

University of Science and Technology of China

Abstract: In this paper, we investigate the geometric fluctuations of principal subspaces for high-dimensional covariance matrices. Specifically, we establish the asymptotic distribution of the sinΘ distance between the sample eigenspace associated with the r_p largest eigenvalues and its population counterpart. Our central limit theorem is derived under notably mild conditions, particularly accommodating a diverging number of spiked eigenvalues r_p, a diverging covariance spectral norm, and the presence of highly heterogeneous spikes. Furthermore, our theoretical results yield deep insights into existing upper bounds in the literature, demonstrating that several known bounds can be strictly sharpened and confirming the minimax optimality of the sample covariance matrix under specific population structures. Various numerical studies further empirically support our theoretical findings.

Stance Drift: How AI-Mediated Communication Distorts Our Message

Lingchong Liu[1],Yanfei Zhou[2],Jacob Bien[2],Y.X. Rachel Wang[3],Lucy Xia[1],Xin Tong[2,4]

1. The Hong Kong University of Science and Technology
2. University of Southern California
3. University of Sydney
4. University of Hong Kong

Abstract： Large language models (LLMs) increasingly mediate human communication, from drafting emails to summarizing scientific reports, yet whether they faithfully preserve a speaker's position remains largely untested. We modeled AI-mediated communication as a two-step pipeline in which one LLM generates an argument from a specified stance and another extracts the stance from that argument. Across 112 debate propositions, 9 mainstream LLMs, and 5 variants, no model preserved the original stance more than 70% of the time. Patterns of failure include polarization, deviation from neutrality, and stance flipping. These results expose a fundamental fidelity gap in AI-mediated communication, with direct implications for journalism, policy deliberation, and most domains where opinion-laden messages pass through language models.

CalCS: Calibrated Cost-Sensitive Classification under Strict Error Constraints

John Park[1],Xin Tong[1],Rachel Wang[2]

1. Hong Kong University
2. The University of Sydney

Abstract： While fundamentally distinct in their objectives, the Neyman-Pearson (NP) and Cost-Sensitive (CS) paradigms both provide essential frameworks for classification. Standard CS methods often rely on empirical error estimates to satisfy user constraints. However, we demonstrate that these plug-in estimators systematically underestimate true population risk. To resolve this, we introduce Calibrated Cost-Sensitive Classification (CalCS), an algorithm that formally bridges the NP and CS paradigms. Given a strict NP error constraint, CalCS leverages exact finite-sample probability bounds to determine associated CS costs that provide the desired control. We show that the algorithm provides finite-sample guarantees while converging to the strict theoretical thresholds. Evaluations on synthetic and real-world datasets demonstrate that CalCS controls targeted error rates, successfully translating mathematical guarantees into robust, practical classifiers.

Computational and Statistical Asymptotic Analysis of the JKO Scheme with Unknown Parameters

Shang Wu

Fudan University

Abstract： The seminal paper of Jordan, Kinderlehrer, and Otto introduced what is now widely known as the JKO scheme, an iterative algorithmic framework for computing distributions. This scheme can be interpreted as a Wasserstein gradient flow and has been successfully applied in machine learning contexts.

In this project, we extend the JKO scheme to accommodate models with unknown parameters. Specifically, we develop statistical methods to estimate these parameters and adapt the JKO scheme to incorporate the estimated values. To analyze the adopted statistical JKO scheme, we establish an asymptotic theory via stochastic partial differential equations that describes its limiting dynamic behavior. Our framework allows both the sample size used in parameter estimation and the number of algorithmic iterations to go to infinity. This study offers a unified framework for joint computational and statistical asymptotic analysis of the statistical JKO scheme. On the computational side, we examine the scheme's dynamic behavior as the number of iterations increases, while on the statistical side, we investigate the large-sample behavior of the resulting distributions computed through the scheme. We conduct numerical simulations to evaluate the finite-sample performance of the proposed methods and the developed asymptotic theory.

Structural Change Detection in Dynamic Systems

Wei Zhang[1],姚方 [2]

1. Fudan University
2. Peking University

Abstract： Structural changes often arise in real-world dynamic systems due to external interventions or environmental shifts, such as policy changes in epidemiology or climate forcing in environmental science. In this paper, we propose a unified framework for detecting and localizing structural changes in dynamic systems governed by ordinary differential equations. Unlike existing methods that assume mean or linear trend changes, our approach accommodates complex, nonlinear dynamics with both stable and diverging trajectories. We develop a new test statistic that combines residual-based discrepancy and normalized parameter contrast, capturing evidence for structural changes from both model fit and parameter shifts. Candidate structural changes are efficiently screened using a multiscale seeded-narrowest-over-threshold algorithm with a data-driven thresholding strategy. To refine selections and control false discoveries, we introduce a false discovery rate control procedure that leverages order-preserved sample splitting and symmetric contrast calibration. Theoretical guarantees are established, including detection consistency, near-minimax localization accuracy, and valid FDR control under weak dependence. Extensive simulations demonstrate superior performance over existing methods in both accuracy and FDR control. Applications to real-world data sets, including COVID-19 dynamics and global temperature trends, highlight the practical relevance and broad applicability of our method.

A Distributional Measure of Heterogeneous Variable Importance

Guorong Dai[1],戴国榕 [1],陈金波 [2]

1. Fudan University
2. University of Pennsylvania

Abstract： When using complex modeling techniques to predict an outcome Y, it is often critical to quantify the contribution of specific covariates Z, commonly termed “variable importance”. In many scientific applications, variable importance exhibits two notable features: (i) heterogeneity with respect to individual characteristics (e.g., age in biomedical research), and (ii) non-transferability between different models, such as conditional mean and quantile models. Existing methods address either (i) or (ii), but not both. To fill this gap, we propose a nonparametric measure that quantifies the full spectrum of Y’s dependence on Z given other baseline covariates, evaluated locally at a fixed point S = s where S denotes the individual characteristics of interest. The measure (a) varies with S to capture individual heterogeneity, and (b) reflects model-agnostic importance with respect to the entire response distribution, rather than a single functional aspect. We construct an intuitive, fully nonparametric estimator based on the K-nearest neighbors algorithm, establish its consistency, and derive its convergence rate. Building on this estimator, we develop a forward selection procedure for identifying covariates locally relevant to Y at S = s, and prove its selection consistency. Simulation studies demonstrate strong finite-sample performance, and two real data applications yield scientifically meaningful findings, underscoring the practical utility of our framework.

Opportunities and Challenges of Sharpness-Aware Training for Overparameterized Neural Networks

Chengli Tan

Northwestern Polytechnical University

Abstract： Recently, a vast array of optimization algorithms has emerged to accelerate training convergence and enhance model generalization. Among them, Sharpness-Aware Minimization (SAM) has gained significant traction because of its remarkable efficacy in improving the generalization of over-parameterized neural networks. However, despite its empirical success, the underlying theoretical mechanisms remain to be fully elucidated. In this talk, I will present some results on this topic, including: 1) investigate why SAM generalize better than SGD from the perspective of algorithmic stability; 2) reveal why SAM calibrate better than SGD through the lens of implicit regularization; 3) introduce an algorithm to accelerate the training speed of SAM by leveraging the local structure of the loss landscape. At last, I will also discuss some limitations of SAM and propose potential strategies to address them.

A Statistical Framework for Alignment with Biased AI Feedback

Zhanrui Cai

The University of Hong Kong

Abstract: Modern alignment pipelines are increasingly replacing expensive human preference labels with evaluations from large language models (LLM-as-Judge). However, AI labels can be systematically biased compared to high-quality human feedback datasets. In this paper, we develop two debiased alignment methods within a general framework that accommodates heterogeneous prompt–response distributions and external human-feedback sources. Debiased Direct Preference Optimization (DDPO) augments standard DPO with a residual-based correction and density-ratio reweighting to mitigate systematic bias, while retaining DPO's computational efficiency. Debiased Identity Preference Optimization (DIPO) directly estimates human preference probabilities without imposing a parametric reward model. We provide theoretical guarantees for both methods: DDPO offers a practical and computationally efficient solution for large-scale alignment, whereas DIPO serves as a robust, statistically optimal alternative that attains the semiparametric efficiency bound. Empirical studies on sentiment generation, summarization, and single-turn dialogue demonstrate that the proposed methods substantially improve alignment efficiency and recover performance close to that of an oracle trained on fully human-labeled data.

Labels or Preferences? Budget-Constrained Learning with Human Judgments over AI-Generated Outputs

Ruijia Wu

Shanghai Jiao Tong University

Abstract: The increasing reliance on human preference feedback to judge AI-generated pseudo labels has created a pressing need for principled, budget-conscious data acquisition strategies. We address the crucial question of how to optimally allocate a fixed annotation budget between ground-truth labels and pairwise preferences in AI. Our solution, grounded in semi-parametric inference, casts the budget allocation problem as a monotone missing data framework. Building on this formulation, we introduce Preference-Calibrated Active Learning (PCAL), a novel method that learns the optimal data acquisition strategy and develops a statistically efficient estimator for functionals of the data distribution. Theoretically, we prove the asymptotic optimality of our PCAL estimator and establish a key robustness guarantee that ensures robust performance even with poorly estimated nuisance models. Our flexible framework applies to a general class of problems, by directly optimizing the estimator's variance instead of requiring a closed-form solution. This work provides a principled and statistically efficient approach for budget-constrained learning in modern AI.

Efficient machine unlearning with minimax optimality

Sai Li[1],Linjun Zhang[2]

1. Tsinghua University
2. Rutgers University

Abstract： In machine learning and artificial intelligence, there is a growing demand for efficient data removal. Regulatory frameworks such as the General Data Protection Regulation grant individuals the right to request the deletion of their personal data from trained models. Furthermore, removing the influence of biased or corrupted data points is essential for model safety and fairness. These requirements have motivated the field of machine unlearning, which aims to eliminate the influence of specific data subsets without the computational cost of full retraining.

In this work, we propose a statistical framework for machine unlearning with convex loss functions and establish theoretical guarantees. For squared loss, especially, we develop Unlearning Least Squares (ULS) and establish its minimax optimality for estimating the model parameter of remaining data when only the pre-trained estimator, forget samples, and a small subsample of the remaining data are available. Our results reveal that the estimation error decomposes into an oracle term and an unlearning cost determined by the forget proportion and the discrepancy between the forget and retained distributions. We further establish asymptotically valid inference procedures without requiring full retraining. Numerical experiments and real-data applications demonstrate that the proposed method achieves performance close to retraining while requiring substantially less data access.

Anti-Concentration Inequalities for the Difference of Maxima of Gaussian Random Vectors

Shuting Shen[1],Alexandre Belloni[2],Ethan Fang[2]

1. National University of Singapore
2. Duke University

Abstract： We derive novel anti-concentration bounds for the difference between the maximal values of two Gaussian random vectors under various settings. Our bounds are dimension-free, scaling with the dimension of the Gaussian vectors only through the smaller expected maximum of the Gaussian subvectors. Meanwhile, our bounds remain valid under degenerate covariance structures, which previous results do not cover. In addition, we show that our conditions are sharp under the homogeneous component-wise variance setting, while we only impose some mild assumptions on the covariance structures under the heterogeneous variance setting. We apply the new anti-concentration bounds to derive the central limit theorem for the maximizers of discrete empirical processes. Finally, we back up our theoretical findings with comprehensive numerical studies.

Active Hypothesis Testing under Computational Budgets

Yin Xia

Fudan University

Abstract： In large-scale hypothesis testing, computing exact p-values or e-values is often resource-intensive, creating a need for budget-aware inferential methods. We propose a general framework for active hypothesis testing that leverages inexpensive auxiliary statistics to allocate a global computational budget. For each hypothesis, our data-adaptive procedure probabilistically decides whether to compute the exact test statistic or a transformed proxy, guaranteeing a valid p-value or e-value while satisfying the exact budget constraint. Theoretical guarantees are established for our constructions, showing that the procedure achieves optimality for e-values and for p-values under independence, and admissibility for p-values under general dependence. Empirical results from simulations and two real-world applications, including a large-scale genome-wide association study (GWAS) and a clinical prediction task leveraging large language models (LLM), demonstrate that our framework improves statistical efficiency under fixed resource limits.

Power Enhancement for Test of Multivariate Independence

许凯 [2],Yeqing Zhou[1],朱利平 [3],李润泽 [4]

1. Tongji University
2. Anhui Normal University
3. Renmin University of China
4. The Pennsylvania State University

Abstract： Testing for independence between two random vectors is a fundamental problem in statistics. It is observed from empirical studies that many existing omnibus consistent tests may not perform well for some strongly nonmonotonic and nonlinear dependence. To get insights into this issue, this paper reveals that a class of existing multivariate independence tests may lose their power due to cancellation of positive and negative terms in dependence metrics. The cancellation leads to the sum of these terms very close to zero. Motivated by this finding, we propose a class of consistent metrics indexed by a positive integer $gamma$ to characterize independence. We further prove that the metrics with even or infinity $gamma$ can effectively avoid the cancellation, and have better powers under the alternatives in which two mean differences offset each other. In practice, it is desirable to target at a wide range of dependence scenarios. Thus, we further advocate to combine the p-values of test statistics with different $gamma$'s. The advantages of the newly proposed tests are illustrated by numerical studies.

Model-X Knockoff Framework for Genome-Wide Survival Association Analysis

Shiyang Ma

Shanghai Jiao Tong University

Abstract： In genome-wide survival association studies, time-to-event (TTE) phenotypes are often underutilized due to the challenges of multiple testing under local linkage disequilibrium (LD) and heavy censoring. We propose Cox-MK, a novel genome-wide survival analysis framework that integrates knockoff statistics with the saddlepoint approximation (SPA), enabling SNP-level false discovery rate (FDR) control in biobank-scale studies. We further introduce SurvKnock, a scalable gene-based framework that leverages model-X knockoffs for conditional survival association testing with rigorous FDR control. SurvKnock employs multiple knockoffs to mitigate LD-induced confounding and integrates frailty-based survival models with SPA, enabling robust inference for both common and rare variants in the presence of sample relatedness and low event rates. Simulation studies and applications to UK Biobank data demonstrate that the proposed methods achieve higher statistical power and well-calibrated FDR control compared with existing approaches, providing effective tools for prioritizing causal variants and genes underlying complex survival phenotypes.

Is the F-test doubly robust?

Lucy Xia

The Hong Kong University of Science and Technology

Abstract： We study the robustness of the F-test in random design linear models, and reach a somewhat nuanced conclusion. On the positive side, one of our main results is that the size of the test is close to its nominal level as soon as either the distribution of the normalised error vector is close to uniform on the unit sphere, or the distribution of the design matrix, after applying any column space-preserving orthogonalisation scheme (e.g.~Gram--Schmidt), is close to a Haar distribution. This provides a sense in which the F-test is doubly robust. Our conclusion is reached by establishing a H\"older continuity property of the Kolmogorov distance between the distribution of the F-statistic and its notional F-distribution under the null. Writing n, p and p_0 for the sample size and the dimensions of the full and null models respectively, we prove that the H\"older exponent is $1/3$ when $\min(n-p,p-p_0) = 1$ and $1/2$ when $\min(n-p,p-p_0) \geq 2$. On the other hand, these exponents are relatively small and cannot be improved in general, revealing that the size of the test may depart from its nominal level quite quickly as we move away from settings where the test is exact. In particular, the regression t-test for the significance of a single predictor corresponds to the case $p - p_0 = 1$, so this test may be especially vulnerable to model misspecification.

High-dimensional Inference for Extreme Value Indices

Liujun Chen

University of Science and Technology of China

Abstract： When applying multivariate extreme value statistics to analyze tail risk in compound events defined by a multivariate random vector, one often assumes that all dimensions share the same extreme value index. While such an assumption can be tested using a Wald-type test, the performance of such a test deteriorates as the dimensionality increases. This paper introduces novel tests for comparing extreme value indices in high-dimensional settings, under both weak and general cross-sectional tail dependence. We establish the asymptotic behavior of the proposed tests. The proposed tests significantly outperform existing methods in high-dimensional scenarios in simulations. We demonstrate real-life applications of the proposed tests for two datasets previously assumed to have identical extreme value indices across all dimensions.

Asymptotic Theory and Penalized Estimation for Signal-plus-Noise Matrices with Heteroskedastic Noise and Weak Signals

Zhixiang Zhang

University of Macau

Abstract： This talk begins with the asymptotic properties of spiked eigenvalues and eigenvectors in signal-plus-noise models when the dimension and size are comparable. Specifically, we assume that the noise has a general covariance matrix that allows for heteroskedasticity, and that the deterministic signal has the same magnitude as the noise, with a rank that may diverge. We then present a new penalized approach to estimate the number of spikes using the explained variance ratio and the rigidity of the non-spiked eigenvalues. This approach has broad applicability beyond the signal-plus-noise model, offering a robust solution for estimating the number of significant components in high-dimensional principal component analysis.

Probabilistic PCA on tensors

Yaoming Zhen

The Chinese University of Hong Kong, Shenzhen

Abstract： In probabilistic principal component analysis (PPCA), an observed vector is modeled as a linear transformation of a low-dimensional Gaussian factor plus isotropic noise. We generalize PPCA to tensors by constraining the loading operator to have Tucker structure, yielding a probabilistic multilinear PCA model that enables uncertainty quantification and naturally accommodates multiple, possibly heterogeneous, tensor observations. We develop the associated theory: We establish identifiability of the loadings and noise variance and show that—unlike in matrix PPCA—the maximum likelihood estimator (MLE) exists even from a single tensor sample. We then study two estimators. First, we consider the MLE and propose an expectation–maximization (EM) algorithm to compute it. Second, exploiting that Tucker maps correspond to rank-one elements after a Kronecker lifting, we design a computationally efficient estimator for which we provide finite-sample guarantees. Together, these results provide a coherent probabilistic framework and practical algorithms for learning from tensor-valued data.

Machine learning methods to predict amyloid positivity using domain scores from cognitive tests

Guogen Shan

University of Florida

Abstract: Amyloid-Beta is the target in many clinical trials for Alzheimer's disease (AD). Preclinical AD patients are heterogeneous with regards to different backgrounds and diagnosis. Accurately predicting Amyloid-Beta status of participants by using machine learning (ML) models based on easily accessible data, could improve the effectiveness of AD clinical trials. We will develop optimal ML models for each subpopulation stratified by sex and disease stages using sub scores from screening neurological tests. Data from the AD Neuroimaging Initiative (ADNI) were used to build the ML models, for three groups: individuals with significant memory concern, early mild cognitive impairment (MCI), and late MCI. Data were further separated into 6 groups by disease stage (3 levels) and sex (2 categories). The outcome was defined as the A status confirmed by the PET imaging, and the features include demographic data, newly identified risk factors, screening tests, and the domain scores from screening tests. Monte Carlo simulation studies were used together with k-fold cross-validation technique to compute model performance metric. We also develop a new feature selection method based on the stochastic ordering to avoiding searching all possible combinations of features. Accuracy of the identified optimal model for SMC male was over 90% by using domain scores, and accuracy for LMCI female was above 86%. Domain scores can improve the ML model prediction as compared to the total scores. Accurate ML prediction models can identify the proper population for AD clinical trials.

Statistical methods for cell-cell interaction studies on spatially resolved transcriptomics

Xiaoyu Song

Duke-NUS Medical School

Abstract: Spatially resolved transcriptomics (SRT) offers unprecedented opportunities to characterize cell–cell interactions (CCI) within intact tissues, yet robust statistical frameworks are critical to uncover true interaction-driven signals. We present three complementary methods addressing this challenge. QuadST provides a powerful framework for single-cell SRT by modeling gene expression across multiple cell–cell distance quantiles to detect interaction-changed genes. It requires no pre-specified interacting cells and is resilient to confounding and measurement error, achieving well-controlled false discovery rates and high statistical power. RECCIPE allows CCI analysis in multi-cell SRT by integrating gene expression, spatial context, and cell-type composition within a multivariate regression framework, yielding accurate inference and novel biological insights. sCCIgen serves as a real-data–based, high-fidelity SRT simulator that generates realistic datasets with tunable CCI patterns under controlled conditions to guide study design and benchmarking. Together, these methods form a rigorous statistical foundation for detecting, validating, and benchmarking CCIs in SRT, advancing both methodological development and scientific discoveries.

Forecasting Respiratory Infectious Diseases with Adaptive and Physics-Informed Machine Learning

Kam Lun Tsang

The University of Hong Kong

Abstract： Accurate short-term forecasting of respiratory infectious diseases is critical for public-health preparedness, but conventional approaches assume stable seasonality and stationary dynamics, assumptions routinely violated by irregular subtropical seasonal patterns, interannual variability in outbreak timing, and the distribution shifts triggered by COVID-19 non-pharmaceutical interventions. We present a programme of machine-learning methods developed to address these challenges across multiple pathogens and regions.

First, we introduced an adaptive-weight ensemble for influenza forecasting in the irregular subtropical context of Hong Kong SAR, China, which outperformed conventional ensembles by dynamically re-weighting component models as seasonal patterns evolve. Second, we extended this concept to respiratory syncytial virus (RSV) by developing Windowed Weighted Average Ensemble (WWAE) and Temporal Simple Average Ensemble (TSAE) methods that allow model selection and weighting to evolve with forecast horizon and temporal distance. Applied to RSV surveillance in Hong Kong SAR (2004–2019) and Japan (2013–2025) with 11 statistical, machine-learning, and deep-learning base models, these adaptive ensembles achieved 31–47% RMSE reductions versus baseline. Third, to tackle the abrupt distribution shifts introduced by the pandemic, we developed a hybrid physics-informed framework (SIRSPF-H) coupling a SIRS compartmental model with a particle filter to estimate real-time transmission dynamics, which are then combined with holiday effects and fed into four deep-learning architectures (GRU, TCN, I-Transformer, N-BEATS). Evaluated on post-pandemic data, SIRSPF-H reduced the weighted interval score by 29–68% across Hong Kong SAR influenza, United States influenza, and Japan RSV.

Together, these studies show that allowing ensembles to adapt temporally and integrating mechanistic constraints with data-driven learning provides a transferable strategy for operational multi-horizon forecasting of respiratory pathogens under irregular and nonstationary dynamics, strengthening real-time surveillance and decision support across diverse epidemiological contexts.

Joint modeling in presence of informative censoring on the retrospective time scale with application to palliative care research

Quran Wu[1],Zhigang Li[2]

1. *Jiangsu Hengrui Pharmaceuticals Co., Ltd.*
2. University of Florida

Abstract： Joint modeling of longitudinal data such as quality of life data and survival data is important for palliative care researchers to draw efficient inferences because it can account for the associations between those two types of data. Modeling quality of life on a retrospective from death time scale is useful for investigators to interpret the analysis results of palliative care studies which have relatively short life expectancies. However, informative censoring remains a complex challenge for modeling quality of life on the retrospective time scale although it has been addressed for joint models on the prospective time scale. To fill this gap, 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.

Differentially private sliced inverse regression in the federated paradigm

Chen Xin

Southern University of Science and Technology

Abstract： Sliced inverse regression (SIR), which includes linear discriminant analysis (LDA) as a special case, is a popular and powerful dimension reduction tool. In this work, we extend SIR to address the challenges of decentralized data, prioritizing privacy and communication efficiency. Our approach, termed as federated sliced inverse regression (FSIR), facilitates distributed computing of the sufficient dimension reduction subspace among multiple clients, solely sharing local estimates to protect sensitive datasets from exposure. To guard against potential adversary attacks, FSIR employs diverse perturbation strategies, including a novel vectorized Gaussian mechanism that guarantees differential privacy at a low cost of statistical accuracy. Additionally, FSIR achieves a tight composition of various privacy mechanisms by adopting a hypothesis testing perspective on differential privacy. It also incorporates a collaborative feature screening procedure, enabling effective handling of high-dimensional client data with varying feature sets. Theoretical properties of FSIR are established for both low-dimensional and high-dimensional settings, supported by extensive numerical experiments and real data analysis.

Adapting to noise tails in private linear regression

Lin Yang

Southwestern University of Finance and Economics

Abstract： Privacy protection has become an increasingly important topic in modern statistical practice. Differential privacy provides a principled framework for protecting sensitive individual information, while robustness is crucial when data exhibit heavy-tailed errors. In this talk, I will discuss differentially private methods for linear regression that combine privacy protection with Huber-type robustification. The talk will cover both low-dimensional and high-dimensional sparse settings, based respectively on noisy clipped gradient descent and noisy iterative hard thresholding. I will discuss how these methods perform under both sub-Gaussian and heavy-tailed errors. In particular, under sub-Gaussian errors, they achieve near-optimal convergence rates while relaxing several assumptions required in earlier work. Under heavy-tailed errors, I will show how the convergence rate depends explicitly on the moment index, privacy parameters, sample size, and intrinsic dimension. The talk will also highlight how these factors influence the choice of robustification parameter and the resulting trade-off among bias, privacy, and robustness.

Federated PCA: Differential Privacy, Algorithms, and Optimality under the Spiked Model

Dong Xia

Hong Kong University of Science and Technology

Abstract： Federated learning (FL) is a distributed learning framework in which a central server communicates with local clients, enabling effective learning while preserving the differential privacy of the clients. Although FL has been widely adopted in industrial and business applications, our understanding of its theoretical limits is still in its infancy. In this talk, we present our results on the classical principal component analysis (PCA) problem in the federated learning setting with heterogeneous privacy constraints. Under mild conditions, our method achieves the minimax-optimal rate under the spiked covariance model.

Site Percolation Network Models for Event-Driven Systems

Yingying Li,Xinghua Zheng

HKUST

Abstract： Conventional network models based on bond percolation (randomness assigned to edges) fail to capture transitivity. We establish a theoretical framework for an alternative class of models based on site percolation (randomness assigned to nodes), which captures transitivity by representing connectivity as the simultaneous activation of nodes, and is therefore suitable for a broad range of event-driven systems. We develop algorithms to consistently estimate this class of network models, covering both pure and mixed membership structures. We also show that the algorithms can be viewed as GMM estimators that achieve semiparametric efficiency. Our framework formalizes and generalizes the models used in recent studies of stock co-jump networks. We further demonstrate the model's practical relevance by recovering latent structures in macroeconomic shocks, firm characteristics co-movements, and co-citation networks.

Cross-Semanic Transfer Learning for High-dimensional Linear Regression

Xuejun Jiang

South University of Science and Technology of China

Abstract： Current transfer learning methods for high-dimensional linear regression assume feature alignment across domains, restricting their applicability to semantically matched features. In many real-world scenarios, however, distinct features in the target and source domains can play similar predictive roles, creating a form of cross-semantic similarity. To leverage this broader transferability, we propose the Cross-Semantic Transfer Learning (CSTL) framework. It captures potential relationships by comparing each target coefficient with all source coefficients through a weighted fusion penalty. The weights are derived from the derivative of the SCAD penalty, effectively approximating an ideal weighting scheme that preserves transferable signals while filtering out source-specific noise. For computational efficiency, we implement CSTL using the Alternating Direction Method of Multipliers (ADMM). Theoretically, we establish that under mild conditions, CSTL achieves the oracle estimator with overwhelming probability. Empirical results from simulations and a real-data application confirm that CSTL outperforms existing methods in both cross-semantic and partial signal similarity settings.

Multi-Source Domain Adaptation via Alignment-Guided Distributionally Robust Learning

Zirui Wang[1],郭子剑[2],王天颖[4],刘默雷[3]

1. Tsinghua University
2. Zhejiang University
3. Peking University
4. Colorado State University

Abstract: We study domain adaptation with labeled data from multiple source environments and unlabeled covariates from a target population. We consider settings where the outcome mechanism is stable across environments after accounting for relevant covariates, but some of these covariates are unobserved. This induces uncertainty about the target population and complicates reliable transfer. To address this challenge, we propose AG-DRoL, an alignment-guided distributionally robust learning framework for multi-source adaptation under unmeasured confounding. The key idea is to use the shared structure across heterogeneous source environments to restrict the class of target distributions compatible with the observed data. This yields a less conservative robust procedure than single-source approaches while retaining protection against latent distributional shift. We establish theoretical guarantees, develop a practical debiased estimator, and show in simulations and a real-data application that AG-DRoL improves worst-case and out-of-sample performance.

Transfer Learning for Spatial Autoregressive Models with Application to U.S. Presidential Election Prediction

Hao Zeng[1],钟威[2],许杏柏[2]

1. Southern University of Science and Technology
2. Xiamen University

Abstract: It is important to incorporate spatial geographic information into U.S. presidential election analysis, especially for swing states. The state-level analysis also faces significant challenges of limited spatial data availability. To address the challenges of spatial dependence and small sample sizes in predicting U.S. presidential election results using spatially dependent data, we propose a novel transfer learning framework within the SAR model, called as tranSAR. Classical SAR model estimation often loses accuracy with small target data samples. Our framework enhances estimation and prediction by leveraging information from similar source data. We introduce a two-stage algorithm, consisting of a transferring stage and a debiasing stage, to estimate parameters and establish theoretical convergence rates for the estimators. Additionally, if the informative source data are unknown, we propose a transferable source detection algorithm using spatial residual bootstrap to maintain spatial dependence and derive its detection consistency. Simulation studies show our algorithm substantially improves the classical two-stage least squares estimator. We demonstrate our method's effectiveness in predicting outcomes in U.S. presidential swing states, where it outperforms traditional methods. In addition, our tranSAR model predicts that the Republican Party would win the 2024 U.S. presidential election.

Federated Learning of Quantile Inference under Local Differential Privacy

Shuyuan Wu,胡祺睿

Shanghai University of Finance and Economics

Abstract： In this paper, we investigate federated learning for quantile inference under local differential privacy (LDP). We propose an estimator based on local stochastic gradient descent (SGD), whose local gradients are perturbed via a randomized mechanism with global parameters, making the procedure tolerant of communication and storage constraints without compromising statistical efficiency. Although the quantile loss and its corresponding gradient do not satisfy standard smoothness conditions typically assumed in existing literature, we establish asymptotic normality for our estimator as well as a functional central limit theorem. The proposed method accommodates data heterogeneity and allows each server to operate with an individual privacy budget. Furthermore, we construct confidence intervals for the target value through a self-normalization approach, thereby circumventing the need to estimate additional nuisance parameters. Extensive numerical experiments and real data application validate the theoretical guarantees of the proposed methodology.

Minimax and Adaptive Covariance Matrix Estimation under Differential Privacy

Yicheng Li

Tsinghua University

Abstract： The covariance matrix plays a fundamental role in the analysis of high-dimensional data. This paper studies minimax and adaptive estimation of high-dimensional bandable covariance matrices under differential privacy constraints. We propose a novel differentially private blockwise tridiagonal estimator that achieves minimax-optimal convergence rates under both the operator norm and the Frobenius norm. In contrast to the non-private setting, the privacy-induced error exhibits a polynomial dependence on the ambient dimension, revealing a substantial additional cost of privacy.

To establish optimality, we develop a new differentially private van Trees inequality and construct carefully designed prior distributions to obtain matching minimax lower bounds. The proposed private van Trees inequality applies more broadly to general private estimation problems and is of independent interest. We further introduce an adaptive estimator that attains the optimal rate up to a logarithmic factor without prior knowledge of the decay parameter, based on a novel hierarchical tridiagonal approach. Numerical experiments corroborate the theoretical results and illustrate the fundamental privacy-accuracy trade-off.

Communication-efficient Distributed Statistical Analysis under Differential Privacy

Haobo Qi

Beijing Normal University

Abstract: Distributed computing is a fundamental paradigm for large-scale statistical analysis, where data are distributed across multiple machines and processed collaboratively. However, the communication between machines introduces significant privacy risks. Differential privacy provides a principled framework for privacy protection, yet its integration into distributed algorithms brings new challenges. In particular, existing one-step approaches fail to achieve global statistical efficiency when the number of machines exceeds the local sample size. To address this issue, we propose a privacy-preserving algorithm based on the Communicationefficient Surrogate Likelihood (CSL) framework proposed by Jordan etal., (2019). We first show that when the local sample size is larger than the number of machines, a one-step estimator remains globally efficient. In the more challenging regime where the number of machines is larger, we show that naive multi-step extensions fail to achieve oracle efficiency due to a trade-off relationship between optimization error reduction and privacy noise inflation. To overcome this, we introduce a novel averagetyped estimator that eliminates this trade-off, thereby achieving global statistical efficiency with guaranteed privacy protection. We further establish its theoretical properties, develop valid inference procedures, and validate our findings through extensive simulation studies and an application to a hospital visit dataset.

Balancing utility and cost in dynamic treatment regimes

Yuqian Zhang

Renmin University of China

Abstract: Dynamic treatment regimes (DTRs) are personalized, adaptive strategies designed to guide the sequential allocation of treatments based on individual characteristics over time. Before each treatment assignment, covariate information is collected to refine treatment decisions and enhance their effectiveness. The more information we gather, the more precise our decisions can be. However, this also leads to higher costs during the data collection phase. In this work, we propose a balanced Q-learning method that strikes a balance between the utility of the DTR and the costs associated with both treatment assignment and covariate assessment. The performance of the proposed method is demonstrated through extensive numerical studies, including simulations and a real-data application to the MIMIC-III database.

Asymptotics of higher criticism via Gaussian approximation

Jingkun Qiu

Peking University

Abstract： Higher criticism is a large-scale testing procedure that can attain the optimal detection boundary for sparse and faint signals. However, there has been a lack of knowledge in most existing works about its asymptotic distribution for more realistic settings other than the independent Gaussian assumption while maintaining the power performance as much as possible. In this talk, we develop a unified framework to analyze the asymptotic distributions of the higher criticism statistic and the more general multi-level thresholding statistic when the individual test statistics are dependent t-statistics under a finite ($2+\delta$)-th moment condition, $0<\delta\leq1$. The key idea is to approximate the global test statistic by the supremum of an empirical process indexed by a normalized class of indicator or thresholding functions, respectively. A new Gaussian approximation theorem for suprema of empirical processes with dependent observations is established to derive the explicit asymptotic distributions.

High-dimensional Clustering and Signal Recovery under Block Signals

Wu Su

Peking University

Abstract： This paper studies computationally efficient methods and their minimax optimality for high-dimensional clustering and signal recovery under block signal structures. We propose two sets of methods, cross-block feature aggregation PCA (CFA-PCA) and moving average PCA (MA-PCA), designed for sparse and dense block signals, respectively. Both methods adaptively utilize block signal structures, applicable to non-Gaussian data with heterogeneous variances and non-diagonal covariance matrices. Specifically, the CFA method utilizes a cross-block statistic to aggregate and select block signals non-parametrically from data with unknown cluster labels. We show that the proposed methods are consistent for both clustering and signal recovery under mild conditions and weaker signal strengths than the existing methods without considering block structures of signals. Furthermore, we derive both statistical and computational minimax lower bounds (SMLB and CMLB) for high-dimensional clustering and signal recovery under block signals, where the CMLBs are restricted to algorithms with polynomial computation complexity. The minimax boundaries partition signals into regions of impossibility and possibility. No algorithm (or no polynomial time algorithm) can achieve consistent clustering or signal recovery if the signals fall into the statistical (or computational) region of impossibility. We show that the proposed CFA-PCA and MA-PCA methods can achieve the CMLBs for the sparse and dense block signal regimes, respectively, indicating the proposed methods are computationally minimax optimal. A tuning parameter selection method is proposed based on post-clustering signal recovery results. Simulation studies are conducted to evaluate the proposed methods. A case study on global temperature change demonstrates their utility in practice.

Semiparametric Sieve Estimation for Survival Data with Two-layer Censoring

Yudong Wang

University of Pennsylvania

Abstract: Disease registry data provide important information on the progression of disease conditions. However, reports of death or drop-out of patients enrolled in the registry are always subject to a noticeable delay. Reporting delays, together with the administrative censoring that arises from a freeze date in data collection, lead to two layers of right censoring in the data. The first layer results from random drop-out and acts on the survival time. The second layer is the administrative censoring, which acts on the sum of the reporting delay and the minimum of the survival time and random drop-out time. The heterogeneities among patients further complicate data analysis. This paper proposes a novel semiparametric sieve method based on phase-type distributions, in which covariates can be readily accommodated by the accelerated failure time model. A well-orchestrated EM algorithm is developed to compute the sieve maximum likelihood estimator. We establish the consistency and rate of convergence of the proposed sieve estimators, as well as the asymptotic normality and semiparametric efficiency of the estimators for the regression parameters. Comprehensive simulations and a real example of lung cancer registry data are used to demonstrate the proposed method. The results reveal substantial biases if reporting delays are overlooked.

Granular Data: Laying a Solid Foundation for High-Quality Datasets, Empowering AI Applications

Yushan Xue

Central University of Finance and Economics

Abstract: Currently, the shortage of high-quality datasets has become a **key bottleneck** restricting the practical application of artificial intelligence. Traditional understanding mostly remains at the abstract technical level, lacking quantitative standards closely tied to business scenarios. As a result, the production, circulation and application of high-quality data are trapped in a dilemma where quality is “indescribable and hard to deliver”. To address this issue, we propose the concept of **granular data**: encapsulating a complete business event into a data component with minimal data integrity, self-descriptiveness and indivisibility, thus reconstructing the logic of data governance from the source. On this basis, we have established a supporting **four-level quality grading scale** (from L0 Initial Level to L3 Enhanced Level), which provides an operable and quantifiable measurement system for data quality. This has achieved a transformation from “post-hoc data cleaning” to “source governance”, and from “extensive management” to “refined quality control”.

Federated Transfer Learning with Differential Privacy

Mengchu Li[1],Ye Tian[2],Yang Feng[3],Yi Yu[4]

1. University of Birmingham
2. Yale University
3. New York University
4. University of Warwick

Abstract： Federated learning has emerged as a powerful framework for analysing distributed data, yet two challenges remain pivotal: heterogeneity across sites and privacy of local data. In this paper, we address both challenges within a federated transfer learning framework, aiming to enhance learning on a target data set by leveraging information from multiple heterogeneous source data sets while adhering to privacy constraints. We rigorously formulate the notion of federated differential privacy, which offers privacy guarantees for each data set without assuming a trusted central server. Under this privacy model, we study four statistical problems: univariate mean estimation, low-dimensional linear regression, high-dimensional linear regression, and M-estimation. By investigating the minimax rates and quantifying the cost of privacy, we show that federated differential privacy is an intermediate privacy model between the well-established local and central models of differential privacy. Our analyses account for data heterogeneity and privacy, highlighting the fundamental costs associated with each factor and the benefits of knowledge transfer in federated learning.

Evaluating LLMs When They Do Not Know the Answer: Statistical Evaluation of Mathematical Reasoning via Comparative Signals

Linjun Zhang

Rutgers University

Abstract： Evaluating mathematical reasoning in LLMs is constrained by limited benchmark sizes and inherent model stochasticity, yielding high-variance accuracy estimates and unstable rankings across platforms. On difficult problems, an LLM may fail to produce a correct final answer, yet still provide reliable pairwise comparison signals indicating which of two candidate solutions is better. We leverage this observation to design a statistically efficient evaluation framework that combines standard labeled outcomes with pairwise comparison signals obtained by having models judge auxiliary reasoning chains. Treating these comparison signals as control variates, we develop a semiparametric estimator based on the efficient influence function (EIF) for the setting where auxiliary reasoning chains are observed. This yields a one-step estimator that achieves the semiparametric efficiency bound, guarantees strict variance reduction over naive sample averaging, and admits asymptotic normality for principled uncertainty quantification. Across simulations, our one-step estimator substantially improves ranking accuracy, with gains increasing as model output noise grows. Experiments on GPQA Diamond, AIME 2025, and GSM8K further demonstrate more precise performance estimation and more reliable model rankings, especially in small-sample regimes where conventional evaluation is pretty unstable. If time permits, we will also talk about how the AI can transform the statistics community in general.

Optimality Theory for Adaptation under Differential Privacy

Lasse Vuursteen

Duke University

Abstract: In classical high dimensional or nonparametric statistics, it frequently occurs that estimators have to adapt to unknown properties of the underlying parameter class or distribution, such as smoothness or sparsity. Under differential privacy constraints, however, adapting to unknown hyperparameters is known to be significantly more challenging, as typical adaptation schemes such as Lepski's method or cross-validation require multiple re-use of the data, which is costly under the differential privacy framework.

In the talk, I will discuss a general optimality theory for adaptation under the federated differential privacy framework, which generalizes local and central differential privacy: data is distributed across many data holders, each imposing a differential privacy constraint. I will present matching upper and lower bounds that precisely quantify the cost of adaptation under federated differential privacy. Specifically, we delineate when adaptation is possible with little to no cost, and when adaptation incurs more significant penalties or is impossible altogether.

LLM-Powered Deep Panel Modeling

Jingyuan Liu

Xiamen University

Abstract: Panel modeling for economic dynamics is crucial for timely and effective policymaking. However, it typically relies only on low-frequency, high-cost surveys and macroeconomic variables, thus often fails to capture rapid market fluctuations and leads to inaccurate predictions. In this paper, we propose a new framework that integrates large language model (LLM) analyses and social media narratives to enhance the prediction power of dynamic panel modeling. Through narrative corpus constructed from social media data, we introduce a prompt-based GPT model and a series of fine-tuned BERT models to generate high-frequency LLM-induced surrogates for the economic indices of interest. A novel joint modeling strategy is then advocated to transfer the information from these surrogates to enhance the prediction power for the targeted economic indices. To solve the joint objectives, we further develop a new deep panel learning procedure with region-wise homogeneity pursuit, which has its own significance in panel data analysis literature. In addition, conformal-based panel prediction intervals are provided to quantify the uncertainty of the LLM-powered prediction. Empirical and theoretical analyses demonstrate that our approach significantly reduces short-term forecasting errors and more effectively captures abrupt inflationary shifts compared to traditional econometric models.

Statistical Inference for Conditional Group Distributionally Robust Optimization with Cross-Entropy Loss

Zijian Guo

Zhejiang University

Abstract: In multi-source learning with discrete labels, distributional heterogeneity across domains poses a central challenge to developing predictive models that transfer reliably to unseen domains. We study multi-source unsupervised domain adaptation, where labeled data are available from multiple source domains and only unlabeled data are observed from the target domain. To address potential distribution shifts, we propose a novel {\bf C}onditional {\bf G}roup {\bf D}istributionally {\bf R}obust {\bf O}ptimization (CG-DRO) framework that learns a classifier by minimizing the worst-case cross-entropy loss over the convex combinations of the conditional outcome distributions from sources domains. We develop an efficient Mirror Prox algorithm for solving the minimax problem and employ a double machine learning procedure to estimate the risk function, ensuring that errors in nuisance estimation contribute only at higher-order rates.

We establish fast statistical convergence rates for the empirical CG-DRO estimator by constructing two surrogate minimax optimization problems that serve as theoretical bridges. A distinguishing challenge for CG-DRO is the emergence of nonstandard asymptotics: the empirical CG-DRO estimator may fail to converge to a standard limiting distribution due to boundary effects and system instability. To address this, we introduce a perturbation-based inference procedure that enables uniformly valid inference, including confidence interval construction and hypothesis testing.

Variational Bayes for high-dimensional structured mixture model

Juan Shen

Fudan University

Abstract： Bayesian methods are widely employed for variable selection; however, the computational complexity associated with Markov Chain Monte Carlo (MCMC) techniques often limits their scalability in high-dimensional contexts. The computation becomes more challenging in mixture models with a substantial number of latent variables. We propose a variational Bayesian (VB) approach for high-dimensional structured mixture models to identify important variables for subgroup analysis. Our method enables efficient and simultaneous variable selection and parameter estimation by approximating the posterior distribution. We establish model selection consistency and derive contraction rates for esti- mation errors, advancing existing VB theoretical results. Additionally, a coordi- nate ascent variational inference algorithm with data augmentation is developed. Numerical studies illustrate that our method achieves accuracy comparable to MCMC while significantly improving computational efficiency. The effectiveness of our method is validated through real-world applications.

Shape-Adaptive Conformal Prediction with Conditional Validity via Minimax Optimization

Haojie Ren

Shanghai Jiao Tong University

Abstract： Achieving valid conditional coverage in conformal prediction is challenging due to the theoretical difficulty of satisfying pointwise constraints in finite samples. Building upon the characterization of conditional coverage through marginal moment restrictions, we introduce Minimax Optimization Predictive Inference (MOPI), a framework that generalizes prior work by optimizing over a flexible class of set-valued mappings rather than calibrating fixed sublevel sets. This minimax formulation effectively circumvents the structural constraints of predefined score functions, achieving superior shape adaptivity while maintaining a principled connection to the minimization of mean squared coverage error. Theoretically, we provide non-asymptotic oracle inequalities and show that the convergence rate of the coverage error attains the optimal order under regular conditions. Our framework also enables valid inference conditional on sensitive attributes that are available during calibration but unobserved at test time. Empirical results on complex, non-standard conditional distributions demonstrate that MOPI produces more efficient prediction sets than existing baselines.

Wasserstein Generative Regression

Shanshan Song

The Chinese University of Hong Kong

Abstract： We propose a new and unified approach for nonparametric regression and conditional distribution learning. Our approach simultaneously estimates a regression function and a conditional generator using a generative learning framework, where a conditional generator is a function that can generate samples from a conditional distribution. The main idea is to estimate a conditional generator satisfying the constraint that it produces a good regression function estimator. We use deep neural networks to model the conditional generator. Our approach can handle problems with multivariate outcomes and covariates, and can be used to construct prediction intervals. We provide theoretical guarantees by deriving non-asymptotic error bounds and the distributional consistency of our approach under suitable assumptions. We perform numerical experiments to demonstrate the effectiveness and superiority of our approach over some existing approaches in various scenarios.

A robust and scalable framework for high-dimensional volatility estimation

Qianqian Zhu

Shanghai University of Finance and Economics

Abstract： This paper introduces a robust and computationally efficient estimation framework for high-dimensional volatility models in the BEKK-ARCH class. The proposed approach employs data truncation to ensure robustness against heavy-tailed distributions and utilizes a regularized least squares method for efficient optimization in high-dimensional settings. This is achieved by leveraging an equivalent VAR representation of the BEKK-ARCH model. Non-asymptotic error bounds are established for the resulting estimators under heavy-tailed regime, and the minimax optimal convergence rate is derived. Moreover, a robust BIC and a Ridge-type estimator are introduced for selecting the model order and the number of BEKK components, respectively, with their selection consistency established under heavy-tailed settings. Simulation studies demonstrate the finite-sample performance of the proposed method, and two empirical applications illustrate its practical utility. The results show that the new framework outperforms existing alternatives in both computational speed and forecasting accuracy.

High-dimensional autoregressive time series modeling for symmetric matrices

Huiling Yuan

East China Normal University

Abstract: Symmetric matrix-valued time series play an important role, especially in economic and finance, and their dimensions can be very large due to the advancement in technology. However, there is still lack of high-dimensional statistical and econometric tools for them. To fill the gap, this paper proposes an autoregressive model with coefficient matrices being in a bilinear form. We further rearrange the coefficient matrices into a tensor, and Tucker decomposition can then be applied, leading to a supervised factor modeling interpretation. As an important application, this paper considers the realized volatility matrices in finance. The high-frequency dynamic model is designed accordingly, and the non-asymptotic properties are established by combining both the error in estimating volatility matrices and that of autoregressive modeling for low-frequency symmetric matrices. Simulation experiments are conducted to evaluate finite-sample performance of the proposed methodology, and its usefulness is further demonstrated by real analysis on the constituent stocks of S&P 500 Index.

High-dimensional Autoregressive Modeling for Time Series Data with Hierarchical Structures

Lan Li,Shibo Yu,Yingzhou Wang,Guodong Li

The University of Hong Kong

Abstract: Modern applications have made ubiquitous high-dimensional data, especially time-dependent data, with more and more complicated structures, and it also has become more frequent to encounter the scenario of hierarchical relationships among variables. However, there is still a lack of supervised learning tool in the literature for them. To fill this gap, we introduce a new model-designing framework, and it then combines with unsupervised factor modeling tools to form an efficient and interpretable autoregressive model for high dimensional time series with hierarchical structures. An ordinary least squares estimation is considered, and its non-asymptotic properties are established. Moreover, we propose an algorithm to search for estimates, and a boosting method is also suggested for hyperparameter selection. Simulation experiments are conducted to evaluate finite-sample performance of the proposed methodology, and its usefulness is demonstrated by an application to the Personality-120 dataset.

Rate-Multiply Robust Estimation of General Treatment Models via Balanced Weighting

Wei Huang[1],Zeqi Wu[2],Meilin Wang[2],Zheng Zhang[2]

1. University of Melbourne
2. Renmin University of China

Abstract： We consider general treatment models identified through weighted (possibly non-smooth) optimization problems, where the weighting function is called the stablised weight. This framework encompasses binary, multi-valued, continuous, and mixed discrete-continuous treatments, and covers a broad class of parameters of interest, including average, quantile, and asymmetric least squares treatment effects.
In recent years, rate-doubly robust estimation of average and quantile treatment effects has received considerable attention, as it can effectively reduce bias arising from nonparametric estimation of nuisance parameters. Many existing doubly robust methods rely on influence functions and require estimating them accurately, which can be challenging in some cases. We show that by exploiting the covariate-balancing property of stabilized weights, together with carefully chosen balancing functions, one can achieve rate double robustness and, more generally, rate-multiple robustness.
We propose a novel balanced augmented weighting method for constructing stabilised weights and estimating general treatment models. The proposed approach can incorporate a wide range of machine learning and deep learning methods for high-dimensional baseline covariates, helping to alleviate the curse of dimensionality while preserving covariate balance through empirical likelihood calibration. This yields debiased and rate-multiply robust estimation of general treatment effects. Under regularity conditions, we show that the proposed estimator is $sqrt{N}$-asymptotically normal and attains the semiparametric efficiency bound. We also develop a weighted bootstrap procedure for statistical inference that avoids direct estimation of efficient influence or score functions.

Perturbed Double Machine Learning: Nonstandard Inference Beyond the Parametric Length

Mengchu Zheng[1],Matteo Bonvini[1],Zijian Guo[2]

1. Rutgers University
2. Zhejiang University

Abstract： We study inference on a low-dimensional functional $_x0008_eta$ in the presence of infinite-dimensional nuisance parameters. Classical inferential methods are typically based on Wald intervals, whose large-sample validity rests on asymptotic negligibility of nuisance error; for example, influence-curve based estimators (Double/Debiased Machine Learning, DML) are asymptotically Gaussian when nuisance estimators converge faster than $n^{-1/4}$. Although such negligibility can hold even in nonparametric classes, it can be restrictive. To relax this requirement, we propose Perturbed Double Machine Learning, which ensures valid inference even when nuisance estimators converge slower than $n^{-1/4}$. Our proposal is to (i) inject randomness into the nuisance estimation step to generate perturbed nuisance models, each yielding an estimate of $_x0008_eta$ and a Wald interval, and (ii) filter out perturbations whose deviations from the original DML estimate exceed a threshold. For Lasso nuisance learners, we show that, with high probability, at least one perturbation yields nuisance estimates sufficiently close to the truth, so the associated estimator of $_x0008_eta$ is close to an oracle with known nuisances. The union of retained intervals delivers valid coverage even when the DML estimator converges slower than $n^{-1/2}$. The framework extends to general machine-learning nuisance learners, and simulations show coverage when state-of-the-art methods fail.

Towards an efficiency theory on parameter manifolds

Lin Liu

Shanghai Jiao Tong University

Abstract: Asymptotic efficiency theory is one of the pillars in the foundations of modern mathematical statistics. Not only does it serve as a rigorous theoretical benchmark for evaluating statistical methods, but it also sheds light on how to develop and unify novel statistical procedures. For example, the calculus of influence functions has led to many important statistical breakthroughs in the past decades. Responding to the pressing challenge of analyzing increasingly complex datasets, particularly those with non-Euclidean/nonlinear structures, many novel statistical models and methods have been proposed in recent years. However, the existing efficiency theory is not always readily applicable to these cases, as the theory was developed, for the most part, under the often neglected premise that both the sample space and the parameter space are normed linear spaces. As a consequence, efficiency results outside normed linear spaces are quite rare and isolated, obtained on a case-by-case basis. This paper aims to develop a more unified asymptotic efficiency theory, allowing the sample space, the parameter space, or both to be Riemannian manifolds satisfying certain regularity conditions. We build a vocabulary that helps translate essential concepts in efficiency theory from normed linear spaces to Riemannian manifolds, such as (locally) regular estimators, differentiable functionals, etc. Efficiency bounds are established under conditions parallel to those for normed linear spaces. We also demonstrate the conceptual advantage of the new framework by applying it to two concrete examples in statistics: the population Frechet mean and the regression coefficient vector of Single-Index Models.

Weighted Youden Index Maximization

Jialiang Li

National University of Singapore

Abstract： In medical research, it is common practice to combine various biomarkers to improve the accuracy of disease diagnosis. The weighted Youden index (WYI), which assigns unequal weights to sensitivity and specificity based on their relative importance, serves as an important and flexible evaluation metric of diagnostic tests. However, no existing methods have been designed specifically to identify the optimal linear combination of biomarkers that maximizes the WYI. In this paper, we propose a novel method to construct an optimal diagnosis score and determine the best cut-off point at the same time. The estimated combination coefficients and cut-off point are shown to have cube root asymptotics, and their joint limiting distribution is established rigorously. Further, the asymptotic normality of the optimal in-sample WYI is established, and out-of-sample inference for score distribution and comparison is investigated.

Provable imitation learning for nuclear fusion control

Wenlong Mou,Xiaofan Xia,Qin Li

University of Toronto

Abstract： Maintaining stability of magnetically confined plasma is a central obstacle to practical nuclear fusion. Modeled as kinetic Vlasov–Poisson equations, the control problem is notably challenging due to non-linearity, sensitivity to initial conditions, and partial observability. Recent development of advanced AI technologies shows promise in control of plasma systems, while theoretical understanding and principled methodologies are still under-explored.

In this talk, we discuss recent advances in machine learning for plasma control. Starting from an expert controller constructed from a fully observed model, we develop algorithms that learn a feedback policy that operates only on experimentally available measurements. We prove that the learned policy stabilizes the plasma dynamics over long time horizons, and provide non-asymptotic sample complexity guarantees for the learning algorithm. The theories demonstrate the advantage of learning-based control in terms of adaptivity to unknown initial conditions and long-term stability. Empirical results on simulated plasma systems also validate the efficacy of our methods in stabilizing plasma over long time horizons.

This work builds a bridge between statistical learning theory and control of complex physical systems, and represents a step toward theoretically grounded, AI-assisted control strategies for fusion energy. Joint work with Xiaofan Xia and Qin Li.

Stein-Encoder: A White-Box Supervised Encoder via Stein Identities in Multi-Modal Studies

Jiarui Zhang

South China University of Technology

Abstract: In multi-modal biomedical research, integrating high-dimensional genomic data with clinical baselines is essential for precision medicine. However, standard deep neural network approaches often entangle these modalities, obscuring the specific predictive impact of genetic features and leading to possibly suboptimal predictive performance. Motivated by the landmark METABRIC cohort primary breast tumors study, we propose the Stein-Encoder, a white-box supervised framework designed to isolate the genetic signal driving clinical outcomes conditional on nuisance covariates. By leveraging Stein's method and residualization techniques, our approach constructs an interpretable single index that summarizes relevant biological heterogeneity while flexibly incorporating clinical factors and can be used to improve downstream prediction. We establish theoretical guarantees for identification, consistency and efficiency improvement. Applied to the METABRIC cohort, the Stein-Encoder outperforms unsupervised benchmarks in predictive accuracy. Crucially, it achieves structural disentanglement by revealing response-specific biological mechanisms: we find that tumor size is driven primarily by mitotic networks, whereas prognostic indices rely on a distinct proliferation-versus-immune axis. This work contributes a unified, computationally efficient framework that bridges statistical rigor with the representational power of neural networks, enabling interpretable, task-specific and efficient compression of multi-modal health data for a wide range of precision medicine applications, beyond biomarker discovery.

Multi-Fidelity Quantile Regression

Yao Zhang

National University of Singapore

Abstract: High-fidelity (HF) data are often expensive to collect and therefore scarce, making conditional quantiles difficult to estimate accurately. We propose a two-stage, model-agnostic method for multi-fidelity quantile regression. The central idea is a local quantile link: at each covariate value, the HF quantile is represented as an LF quantile evaluated at a covariate-dependent level. This reformulation reduces the problem to estimating the level function, which can be smoother than the HF quantile itself when the LF and HF conditional distributions have similar shapes. We also study the complementary regime in which this advantage weakens and introduce a correction step to improve robustness. Our theory characterizes when the proposed estimator converges faster than direct quantile regression using HF data alone and when the correction step provides further improvement. Experiments on synthetic and real data show that our method yields more accurate quantile estimates and tighter conformal prediction intervals.

All-in-One Toolkit for Biobank-Scale Whole-Genome Sequencing Data Management and Analysis

Zilin Li

Northeast Normal University

Abstract: Biobank-scale Whole-Genome Sequencing (WGS) studies are increasingly pivotal in unraveling the genetic bases of diverse health outcomes. However, managing and analyzing these datasets' sheer volume and complexity presents significant challenges. We propose vcf2agds, an all-in-one toolkit that efficiently converts WGS data from Variant Call Format (VCF) format to the annotated Genomic Data Structure (aGDS) format, significantly reducing data size while supporting seamless genomic and functional data integration for comprehensive genetic analyses. Additionally, STAARpipeline equipped with the aGDS files enabled scalable, comprehensive and functionally informed WGS analysis, facilitating the detection of common and rare coding and noncoding phenotype-genotype associations. We applied the STAARpipeline to analyze Alzheimer disease (AD) in 459,216 samples from the UK Biobank. All analyses scale well in computation time and memory. We discover several potentially new significant associations with AD. As WGS datasets continue to expand in size and complexity, our proposed tools will be increasingly useful for unlocking the full potential of genomic research.

A Powerful Transformation of Quantitative Responses for Biobank-Scale Association Studies

Yaowu Liu

Southwestern University of Finance and Economics

Abstract: In linear regression models with non-Gaussian errors, transformations of the response variable are widely used in a broad range of applications. Motivated by various genetic association studies, transformation methods for hypothesis testing have received substantial interest. In recent years, the rise of biobank-scale genetic studies, which feature a vast number of participants that could be around half a million, spurred the need for new transformation methods that are both powerful for detecting weak genetic signals and computationally efficient for large-scale data. In this work, we propose a novel transformation method that leverages the information of the error density. This transformation leads to locally most powerful tests and therefore has strong power for detecting weak signals. To make the computation scalable to biobank-scale studies, we harnessed the nature of weak genetic signals and proposed a consistent and computationally efficient estimator of the transformation function. Through extensive simulations and a gene-based analysis of spirometry traits from the UK Biobank, we validate that our approach maintains stringent control over Type I error rates and significantly enhances statistical power over existing methods.

A Unified Mendelian Randomization Framework for Identifying Causal Risk Factors: Accounting for Measured Covariates and Unmeasured Confounders

Xianghong Hu

Shenzhen University

Abstract：TBD

Bridging unpaired single-cell multimodal data for integrative analyses with SuperMap

Jingsi Ming

East China Normal University

Abstract： Current single-cell profiling technologies enable the capture of multiple cellular modalities, providing valuable insights into complex biological systems. While a substantial amount of single-cell multimodal data has been generated and accumulated, most of these datasets are unpaired, characterized by distinct feature spaces and a lack of cell-wise correspondence. The absence of explicit linkages between modalities poses a fundamental challenge for data integration and interpretation. To address this, we introduce SuperMap, a statistical learning method designed for the integrative analyses of unpaired multimodal data. SuperMap directly learns cross-modal mappings from unpaired data to effectively bridge and link different modalities, facilitating a variety of downstream analysis tasks. Comprehensive benchmarking and real-world applications demonstrate the superior performance of SuperMap in enhancing cell-type identification, improving diagonal integration, enabling regulatory analysis, and revealing epigenomic priming events to specify cell differentiation directions for trajectory inference.

Deciphering microbial community dynamics using cross-sectional data-informed NeuralODE

Tao Wang

Shanghai Jiao Tong University

Abstract: Understanding the ecological mechanisms of host-associated microbial ecosystems typically relies on either cross-sectional or time-series data. Cross-sectional analyses are limited in their ability to assess intervention effects, whereas time-series models require dense and informative sampling that is often impractical. In this talk, we present an enhanced Neural Ordinary Differential Equations (NeuralODE) framework that, for the first time, integrates cross-sectional data into the dynamic modeling of sparse and weakly informative temporal data. We further introduce a dynamic keystoneness metric to comprehensively quantify species importance over time. Across simulated and real-data benchmarks, incorporating cross-sectional data improved performance over competing methods, particularly in data-scarce settings. Moreover, biological validation demonstrated that the framework recovers experimentally supported interactions and prioritizes identified influential species. Together, these results establish our method as a reliable framework for mechanistic modeling of microbial ecosystems, offering new insights into their dynamic behavior.

Constructive Instrumental Variable Identification and Inference with Many Weak Interaction Moments

Zhonghua Liu

Columbia University

Abstract: Instrumental variable methods are widely used for causal inference, but identification becomes especially challenging when instruments are weak and potentially invalid. These challenges are particularly pronounced in Mendelian randomization, where genetic variants serve as instruments and violations of exclusion restriction or independence assumptions are common. We propose MAGIC, a constructive and assumption-lean framework that achieves identification even when all candidate instruments may be invalid. The method exploits pairwise and higher-order interactions among mutually independent instruments to construct moment conditions orthogonal to both unmeasured confounding and direct effects under a linear structural model. The resulting estimation problem involves many potentially weak interaction moments with unknown nuisance parameters. We develop a semiparametric generalized method of moments estimator and introduce a global Neyman orthogonality condition to ensure robustness of both the moment function and its derivative to nuisance estimation under many weak moment asymptotics. We establish consistency and asymptotic normality when the number of moments diverges with sample size and characterize the semiparametric efficiency bound under fixed dimension. Simulations and an application to UK Biobank data illustrate the method.

Words matter: Multimodal Suicide Risk Prediction from Veterans Health Administration Clinical Notes

Jiang Gui

Dartmouth College

Abstract: In this talk, we demonstrate that integrating unstructured clinical narratives with structured electronic health record (EHR) data enhances suicide risk prediction for U.S. Veterans, outperforming models that rely on structured data alone. By analyzing a retrospective matched case-control cohort of 4,584 Veterans who died by suicide and 22,657 controls, we compared traditional count-based text features against pretrained contextual large language model (LLM) embeddings, such as Clinical Longformer and BioClinicalBERT. We found that while Adaptive Mixture Categorization (AMC) improves the utility of skewed linguistic data, contextual LLM embeddings consistently provide comparable or superior predictive power, particularly within low- and moderate-risk tiers where structured indicators may be less obvious. Our multimodal approach, which integrated 66 structured patient characteristics with text features, yielded substantial performance gains, increasing AUROC by approximately 0.07–0.11 across various risk tiers and time windows. Furthermore, our temporal analysis revealed that while long-term data (270 days) is most informative for low-risk patients, short-term windows (<30 days) are critical for high-risk individuals. Using SHAP-based interpretability and topic modeling, we identified clinically coherent themes that shift semantically as risk increases, providing a context-aware framework for improving suicide prevention efforts within the Veterans Health Administration.

Distributed Learning with Heterogeneity and Asynchrony: Representative Learning

Keren Li

University of Alabama at Birmingham

Abstract: Distributed machine learning faces fundamental challenges arising from data heterogeneity and asynchronous communication, under which gradient- or model-based information exchange becomes unstable, model-dependent, and difficult to interpret. This talk presents Representative Learning (RepL), a unified framework that replaces model-derived summaries with structured pseudo-data, called representatives, that preserve the original data format while encoding key statistical properties of local datasets.

We systematically develop representative constructions across model classes. The Mean Representative (MR) provides a model-agnostic baseline through moment matching, while the Score-Matching Representative (SMR) extends this idea to generalized linear models by aligning local score functions. To address nonlinearity and instability under heterogeneity, we introduce the Transformed Mean Representative (TMR), which incorporates link-function transformations, and the Anchored Score-Matching Representative (Anchored-SMR), which enforces identifiability and stabilizes representative construction through anchored score equations.

The framework is further extended to smooth and non-smooth objectives, and to decentralized and asynchronous systems where communication is delayed or partial. Theoretical results establish convergence under representative approximation error and bounded delay, while empirical studies demonstrate that RepL achieves stable and accurate learning across heterogeneous regimes where classical federated approaches deteriorate.

These results position RepL as a scalable, interpretable, and model-agnostic alternative for distributed learning under heterogeneity and asynchrony.

Unified Conformalized Multiple Testing with Full Data Efficiency

Xiaoyang Wu

Nankai University

Abstract： Conformalized multiple testing offers a model-free way to control predictive uncertainty in decision-making. Existing methods typically use only part of the available data to build score functions tailored to specific settings. We propose a unified framework that puts data utilization at the center: it uses all available data—null, alternative, and unlabeled—to construct scores and calibrate p-values through a full permutation strategy. This unified use of all available data significantly improves power by enhancing non-conformity score quality and maximizing calibration set size while rigorously controlling the false discovery rate. Crucially, our framework provides a systematic design principle for conformal testing and enables automatic selection of the best conformal procedure among candidates without extra data splitting. Extensive numerical experiments demonstrate that our enhanced methods deliver superior efficiency and adaptability across diverse scenarios.

Gold after Randomization: Model-X Split Knockoffs for FDR Control in Transformation Selection

Yuan Yao

The Hong Kong University of Science and Technology

Abstract： Controlling the False Discovery Rate (FDR) in transformation selection has wide applications including neuroimaging studies of Alzheimer's Disease, learning human preference from pairwise comparisons, identifying causal invariant features from spurious ones in multiple environments, and trend filtering in economics. While the recent Split Knockoff method provides finite-sample FDR control in trasnsformation selection, it is limited to fixed designs with linear models. Extending this framework to random designs with a broader class of nonlinear models needs to reconcile a random covariate design with a deterministic linear transformation. To address this challenge, we propose a randomized Split Knockoff method, inspired by the classical Model-X Knockoff and generalizing it to transformation selection, that achieves robust FDR control by incorporating a randomized mapping from the transformation's subspace constraint to a neighborhood in a lifted parameter space. For the important case of pairwise comparisons, we further develop an augmented bootstrap approach that constructs knockoffs without explicit knowledge of marginal covariate distributions. By leveraging the randomization, our method may exhibit statistical power at least equivalent to—and often superior to—the standard Model-X Knockoff in some senarios when both are applicable. We demonstrate the robust FDR control and improved power by both simulations and several real-world applications.

Multiple Testing Meets Data Visualization: A Modern Perspective on Boxplots and Bagplots

Bowen Gang

Fudan University

Abstract: This report develops a unifying multiple testing framework that recasts boxplot outlier detection as graphical hypothesis testing, fundamentally addressing the failure of fixed-threshold rules at large sample sizes. Based on this framework, we introduce adaptive variants controlling FWER, PFER, and FDR, enabling data-dependent fence construction. For bivariate data, we propose the "bag-and-whisker plot," replacing the unstable convex hull with a direct statistical fence visualization. These contributions transform classical exploratory tools into statistically rigorous, visually reliable methods suitable for modern large-scale data analysis.

A New Approach to Conformalized Model Selection for Out-of Distribution Testing

Zinan Zhao

Zhejiang University

Abstract: This talk addresses the challenge of structured out-of-distribution (OOD) testing in high-stakes machine learning (ML) applications. Traditional conformal methods rely on the strict joint exchangeability assumption, render_x0002_ing them unsuitable for data-rich scenarios where valuable auxiliary information– such as spatiotemporal or grouping structures – is available. To overcome this limitation, we present the structure-adaptive conformal q-value (SCQ), a novel significance index that integrates individual test evidence with structural pat_x0002_terns. Additionally, our work develops the pseudo-score-guided transductive automated model selection (P-TAMS) algorithm, which adapts conformalized model selection techniques to leverage structural insights and optimize OOD testing performance across a toolbox of candidate ML models. Together, SCQ and P-TAMS form a unified and flexible framework based on the weaker as_x0002_sumption of pairwise exchangeability, offering guaranteed error rate control, en_x0002_hanced statistical power, and improved interpretability in detecting structured anomalies.

Policy Averaging for Stochastic Decision Problems: Theory and an Application to the Newsvendor Problem

Xiangyu Cui

Shanghai University of Finance and Economics

Abstract： We propose a Policy Averaging Approach (PAA) for stochastic decision problems that combines multiple candidate policies into a single data-driven decision rule. The main idea is to exploit diversification across policies, in the spirit of model averaging and risk diversification, so as to improve robustness, stability, and decision quality under uncertainty. Rather than relying on a single model specification or a single estimated policy, PAA aggregates information from competing policies and uses cross-validation to select and optimize averaging weights. Our main contribution is methodological and general. We formulate PAA for a broad class of stochastic decision problems and provide theoretical analysis to shed light on its effectiveness. The results suggest that policy averaging can improve the stability of data-driven decisions, reduce sensitivity to model misspecification, and enhance out-of-sample performance relative to individual candidate policies. We use the newsvendor problem as an illustrative application. In that setting, existing approaches often depend on restrictive distributional assumptions, specific feature-based demand models, or highly adaptive data-driven rules that may be unstable or prone to overfitting. PAA provides a unified framework that synthesizes such competing approaches while retaining interpretability and empirical flexibility. Through theoretical analysis, simulation experiments, and an empirical study, we show that PAA achieves lower expected cost, more stable performance, and better justified recommendations than benchmark methods.

Distributed propensity model averaging for large-scale data with nonignorable nonresponse

Fang Fang

East China Normal University

Abstract： The proliferation of large-scale data has stimulated extensive research on distributed statistical methods. However, relatively little attention has been paid to distributed methods for missing data. Existing approaches are mainly developed under the missing at random assumption and may suffer bias when the missingness mechanism is missing not at random. Moreover, most model averaging methods for missing data focus on averaging outcome regression models while assuming a correctly specified propensity model. In this paper, we propose a distributed propensity model averaging method for large-scale data with nonignorable nonresponse. We first aggregate the parameters for the propensity model on each machine, and subsequently construct a divide-and-conquer-type model averaging estimator for distributed data. Under mild regularity conditions, we prove that the averaged weights converge in L2 to the theoretically optimal weights minimizing the risk and further derive the L2 convergence of the resulting estimator when there are correct candidate models. Asymptotic optimality is also established even when all candidate models are misspecified. Simulation studies and a real data analysis demonstrate the empirical performance of the proposed method.

Double Descent and Emergence in Multi-model Prediction

Dandan Jiang

Xi'an Jiaotong University

Abstract: This talk proposes a novel model averaging framework designed for high- dimensional data where the number of predictors p is comparable to the sample size n. The core innovation lies in optimizing model weights based on the asymptotic predictive risk, effectively overcoming the limitations of traditional training-error-based criteria. Theoretically, we derive the limiting expression for out-of-sample predictive risk under a nested model setting, which formally characterizes the double-descent, the emergent phase transition at the interpolation boundary, and the smoothing effect induced by the averaging process. Methodologically, we introduce a concrete method, Large Model Averaging (LaMA), which incorporates both the in-sample bias and the predictive variance into the criterion to simultaneously optimize the fitting accuracy and generalization, stabilized by l_2-regularization. Unlike existing approaches that are restricted to $p \ll n$, our framework performs reliably across a wide range of p/n ratios, ensuring both stable optimization and superior generalization by explicitly accounting for the high-dimensional risk landscape.

Semi-supervised learning using copula-based regression and model averaging

Dalei Yu

Xi'an Jiaotong University

Abstract: The available data in semi-supervised learning usually consists of relatively small sized labeled data and much larger sized unlabeled data. How to effectively exploit unlabeled data is the key issue. In this paper, we write the regression function in the form of a copula and marginal distributions, and the unlabeled data can be exploited to improve the estimation of the marginal distributions. The predictions based on different copulas are weighted, where the weights are obtained by minimizing an asymptotic unbiased estimator of the prediction loss. Error-ambiguity decomposition of the prediction loss is performed such that unlabeled data can be exploited to improve the prediction loss estimation. We demonstrate the asymptotic normality and the nonasymptotic error bounds of copula parameters and regression function estimators of the candidate models under the semi-supervised framework, as well as the asymptotic optimality and weight consistency of the model averaging estimator. By incorporating unlabeled data into the candidate model estimation process, one can achieve a sharper error bound in comparison to the supervised counterpart. In addition, our model averaging estimator achieves faster convergence rates of asymptotic optimality and weight consistency than the supervised counterpart. Extensive simulation experiments and the California housing dataset demonstrate the effectiveness of the proposed method.

Neural Wasserstein Two-Sample Tests

Zhenhua Lin[1],Zhenhua Lin[2]

1. Xi'an Jiaotong University
2. National University of Singapore

Abstract： The two-sample homogeneity testing problem is fundamental in statistics and becomes particularly challenging in high dimensions, where classical tests can suffer substantial power loss. We develop a learning-assisted procedure based on the projection Wasserstein distance. The method is motivated by the observation that there often exists a low-dimensional projection under which the two high-dimensional distributions differ. In practice, we learn the projection directions via manifold optimization and a witness function using deep neural networks. To adapt to unknown projection dimensions and sparsity levels, we aggregate a collection of candidate statistics through a max-type construction, avoiding explicit tuning while potentially improving power. We establish the validity and consistency of the proposed test and prove a Berry-Esseen type bound for the Gaussian approximation. In particular, under the null hypothesis, the aggregated statistic converges to the absolute maximum of a standard Gaussian vector, yielding an asymptotically pivotal (distribution-free) calibration that bypasses resampling. Simulation studies and a real-data example demonstrate the strong finite-sample performance of the proposed method.

Ball Impurity: Measuring Heterogeneity in General Metric Spaces

Wenliang Pan

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Abstract： Data in various domains, such as neuroimaging and network data analysis, often come in complex forms without possessing a Hilbert structure. The complexity necessitates innovative approaches for effective analysis. We propose a novel measure of heterogeneity, ball impurity, which is designed to work with complex non-Euclidean objects. Our approach extends the notion of impurity to general metric spaces, providing a versatile tool for feature selection and tree models. The ball impurity measure exhibits desirable properties, such as the triangular inequality, and is computationally tractable, enhancing its practicality and usefulness. Extensive experiments on synthetic data and real data from the UK Biobank validate the efficacy of our approach in capturing data heterogeneity. Remarkably, our results compare favorably with state-of-the-art methods in metric spaces, highlighting the potential of ball impurity as a valuable tool for addressing complex data analysis tasks.

Semi-Supervised Generative Learning via Latent Space Distribution Matching

Long Feng

University of Hong Kong

Abstract: Conditional generative learning plays a vital role in modern machine learning and supports a wide range of applications, including language modeling and image generation. However, most existing methods depend heavily on paired data, which is often costly to obtain. This paper explores conditional generative learning within a semi-supervised framework, leveraging both labeled and unlabeled data to enhance sampling from conditional distributions. We introduce Latent Space Distribution Matching (LSDM), a theoretically grounded semi-supervised approach characterized by two main innovations: (1) a novel objective that provides an upper bound on the true conditional distribution and enables decoupled unsupervised and supervised training phases; and (2) a flexible optimization strategy that accommodates various divergence measures. LSDM employs a two-stage training process, first learning latent representations from all available data, and then matching conditional distributions within this latent space. Our method naturally unifies and extends concepts from pre-training-based methods. Theoretically, we prove that LSDM benefits from both labeled and unlabeled data, enabling the learning of data manifold structure at a rate determined by the total sample size. Experiments on two real-world image tasks—image inpainting and image super-resolution—demonstrate the effectiveness of LSDM and highlight the value of incorporating unlabeled data.

Factor-augmented clustering tree for time series

Ting Li

Southern University of Science and Technology

Abstract: Clustering time series is essential for uncovering subgroups and patterns in large temporal datasets, but existing methods often overlook covariate information and struggle with pervasive common factors. We address these challenges by proposing the novel Covariate-driven Group Factor Model, which links covariates to heterogeneous factor structures, and by introducing the Factor-Augmented Clustering Tree (FACT) algorithm to fit this model. FACT partitions data by distinguishing local from common factors using a new splitting criterion and robust stopping rules. We establish theoretical consistency for group recovery in high dimensions and provide a scalable, interpretable algorithm. Applied to Chinese air quality data, our method reveals geographically coherent regions aligned with major climatological boundaries.

SMODER: Spatial Multi-Omics Deconvolution Anchored by RNA

Wei Liu

Sichuan University

Abstract： Spatial technologies that jointly profile RNA and additional molecular layers offer a comprehensive view of tissue organization, but limited spatial resolution results in measurements that reflect mixtures of cell types. While RNA-based deconvolution methods are well developed, it remains unclear how to incorporate co-profiled non-transcriptomic modalities without compromising cell-type identifiability, as reliable references are largely restricted to RNA. Here we present SMODER, an RNA-anchored framework based on graph neural networks for spatial multi-omics deconvolution that treats transcriptomes as the primary determinant of cell identity while leveraging additional modalities to refine spatial inference. By integrating heterogeneous omics layers in a constrained manner, SMODER preserves RNA-driven cell-type resolution and enables robust estimation of cell-type compositions across spatial locations. Across simulations and multiple spatial multi-omics datasets spanning protein, chromatin accessibility and histone modification assays, SMODER consistently outperforms existing methods in both accuracy and spatial coherence. Beyond deconvolution, SMODER denoises multi-omic signals and enables cell-type-aware analysis of regulatory programs. Application to spatial CUT\&Tag–RNA-seq and RNA–ATAC-seq data reveals spatially resolved epigenetic repression and cis-regulatory relationships linked to gene expression. Together, SMODER establishes a principled framework for RNA-guided integration of spatial multi-omics data, enabling analyses that are not accessible with RNA-only or symmetrically integrated approaches.

Nonparametric Inference on Unlabeled Histograms with Application to Generative Uncertainty Evaluation

Pengkun Yang

Tsinghua University

Abstract： Statistical inference on histograms and frequency counts plays a central role in categorical data analysis. Moving beyond classical methods that directly analyze labeled frequencies, we introduce a framework that models the multiset of unlabeled histograms via a mixture distribution to better capture unseen domain elements in large-alphabet regime. We study the nonparametric maximum likelihood estimator (NPMLE) under this framework, and establish its optimal convergence rate under the Poisson setting. The NPMLE also immediately yields flexible and efficient plug-in estimators for functional estimation problems, where a localized variant further achieves the optimal sample complexity for a wide range of symmetric functionals. Extensive experiments on synthetic, real-world datasets, and large language models highlight the practical benefits of the proposed method. I will also discuss the extensions using structural reasoning and prompt perturbations.

Provable RLHF: A Consistent Framework for Offline Reward Learning and Value Function Learning

Lican Kang

Wuhan University

Abstract： Reinforcement Learning from Human Feedback (RLHF) has become a widely adopted paradigm for aligning AI systems with human preferences, particularly in tasks where explicit reward specification is infeasible. However, developing a unified framework that effectively integrates reward learning and policy learning in RLHF, while providing theoretical guarantees, remains a significant challenge. This work proposes a provably consistent framework for offline RLHF that jointly recovers the underlying reward function and estimates the optimal action-value function from batch data. The framework proceeds in two stages: (1) reward recovery by minimizing the Bellman residual, where a conditional Schrodinger-Follmer flow is employed as a generative model for the transition dynamics, and estimation of the behavior policy's action-value function via maximum likelihood under a softmax policy model; and (2) optimal value function estimation by minimizing a minimax Bellman optimality error under the learned reward. We establish non-asymptotic statistical guarantees for each stage of the framework, including convergence rates for conditional generative modeling, reward recovery, and value estimation. Our results provide the unified theory for RLHF with provable error bounds, bridging the gap between empirical practice and theoretical understanding in RLHF.

Semi-supervised learning in high-dimensional linear regression

Ling Zhou

Southwestern University of Finance and Economics

Abstract： Semi-supervised learning (SSL) has attracted considerable interest for its ability to leverage a large unlabeled sample to improve estimation and prediction beyond what supervised learning (SL) can achieve using labeled data alone. In this paper, we establish minimax learning rates for both estimation and prediction in high-dimensional linear regression under the SSL paradigm. Motivated by the fact that unlabeled data are informative when the covariates exhibit structure aligned with the regression coefficients, we conduct a systematic study across three representative settings: eigen-decay model, spiked model, and factor model. Our theoretical analysis yields several notable findings: (a) in challenging regimes where SL exhibits poor performance, such as those characterized by weak signals or high dimensionality, SSL attains strictly faster convergence rates, with the semi-supervised acceleration ratio scaling proportionally to the ratio of unlabeled to labeled sample sizes for estimation; and (b) SSL methods based on feature extraction achieve minimax optimal estimation rates under substantially broader conditions, whereas imputation-based SSL approaches attain optimality only when the covariate dimension is smaller than the labeled sample size; and (c) for prediction, feature extraction methods uniformly outperform imputation-based approaches. Our theoretical findings are further corroborated by extensive numerical experiments.

Inference-Time Alignment for Diffusion Models via Variationally Stable Doob's Matching

Yuling Jiao

Wuhan University

Abstract: Inference-time alignment for diffusion models aims to adapt a pre-trained diffusion model toward a target distribution without retraining the base score network, thereby preserving the generative capacity of the base model while enforcing desired properties at the inference time. A central mechanism for achieving such alignment is guidance, which modifies the sampling dynamics through an additional drift term. In this work, we introduce Doob's matching, a novel framework for guidance estimation grounded in Doob's -transform. Our approach formulates guidance as the gradient of logarithm of an underlying Doob's -function and employs gradient-penalized regression to simultaneously estimate both the -function and its gradient, resulting in a consistent estimator of the guidance. Theoretically, we establish non-asymptotic convergence rates for the estimated guidance. Moreover, we analyze the resulting controllable diffusion processes and prove non-asymptotic convergence guarantees for the generated distributions in the 2-Wasserstein distance.

Heavy-tailed Information-Theoretic Generalization Bounds with Applications to LLM Safety Alignment

Huiming Zhang[1],李秉翰[1],田万[2],孙强[3]

1. Beihang University
2. Peking University
3. University of Toronto

Abstract: Classical information-theoretic generalization bounds, which link generalization error to the mutual information between an algorithm's input and output, typically rely on sub-Gaussian assumptions or finite moment generating functions (MGFs). However, these assumptions are often violated in heavy-tailed scenarios, such as adversarial training, reinforcement learning with rare high-reward events, and financial modeling. In this work, we bridge this gap by establishing a comprehensive framework for generalization under heavy-tailed sub-Weibull regimes. We demonstrate that standard KL divergence bounds are vacuous in these settings due to the unboundedness of extreme events. To overcome this, we introduce a novel decorrelation lemma based on Rényi divergence and a generalized, MGF-free Young-type inequality. By combining these tools with a refined chaining technique on the space of measures, we derive Dudley-type generalization bounds that explicitly depend on the tail parameter and the Rényi information. Additionally, we establish new maximal inequalities and information-theoretic generalization bounds and stochastic gradient Langevin dynamics (SGLD),under the assumption that the loss functions exhibit heavy-tailed sub-Weibull behavior. We apply our theory to large language models (LLMs) in the context of reward hacking within Reinforcement Learning from Human Feedback (RLHF). We show that Rényi-regularized alignment provides finite reward guarantees and ensures that best-of-N policies remain well-controlled, thereby mitigating the catastrophic Goodhart effects where standard KL-regularization fails.

Symmetry in Deep Neural Networks and Implications to Learning

Guohao Shen

The Hong Kong Polytechnic University

Abstract： Overparameterized deep neural networks achieve remarkable generalization despite their massive parameter counts, challenging classical learning theory. This talk explores this phenomenon through the lens of parameter space symmetries, including scaling, sign-flip, and permutation invariances that lead to functional equivalence. By quotienting out these inherent redundancies, we construct an "effective parameter space" and derive a drastically tighter upper bound for the network's covering number, mathematically reducing the theoretical complexity by a factorial factor of the hidden layer widths. Furthermore, we investigate how these symmetric geometries shape a highly connected loss landscape that naturally facilitates gradient-based optimization. Finally, we will introduce practical, symmetry-aware techniques and aligned model averaging, demonstrating how leveraging these invariances can directly accelerate training and enhance distributed learning efficiency.

A novel statistical approach to analyze image classification

Juntong Chen[1],Sophie Langer[2],Johannes Schmidt-Hieber[3]

1. Xiamen University
2. Ruhr University Bochum
3. University of Twente

Abstract： The recent statistical theory of neural networks focuses on nonparametric denoising problems that treat randomness as additive noise. Variability in image classification datasets does, however, not originate from additive noise but from variation of the shape and other characteristics of the same object across different images. To address this problem, we introduce a tractable model for supervised image classification. While from the function estimation point of view, every pixel in an image is a variable, and large images lead to high-dimensional function recovery tasks suffering from the curse of dimensionality, increasing the number of pixels in the proposed image deformation model enhances the image resolution and makes the object classification problem easier. We introduce and theoretically analyze three approaches. Two methods combine image alignment with a one-nearest neighbor classifier. Under a separation condition, it is shown that perfect classification is possible. The third method fits a convolutional neural network (CNN) to the data. We derive a rate for the misclassification error that depends on the sample size and the complexity of the deformation class. An empirical study corroborates the theoretical findings.

Towards 3D Content Generation and Understanding via Pretrained Generative Priors

Lu Sheng

Beihang University

Abstract：TBD

Realistic Human Interaction in Virtual Environments

Yudong Guo

University of Science and Technology of China

Abstract： Realistic humans and their interactions with surrounding environments are fundamental building blocks of any interactive world model. In this talk, I will present our recent progress along two closely connected fronts. The first focuses on realistic human modeling and generation, aiming at digital humans with high visual fidelity, expressive dynamics, and fine-grained controllability. The second extends from the human body to the human–scene relationship, targeting physically plausible and semantically coherent interactions between digital humans and 3D environments. I will discuss the key challenges behind each direction and share our thoughts on unifying these components to support the next generation of embodied, interactive virtual worlds.

Visual Affective-Cognitive Perception and Modeling

Zhiwen Shao

China University of Mining and Technology

Abstract： Visual affective-cognitive perception and modeling is a core technical support for the construction of embodied intelligence and the embodied world model. Embodied intelligence requires intelligent agents to integrate visual perception, affective understanding, and environmental modeling to achieve natural interaction with humans and adaptive response to complex environments. However, existing technologies still face key challenges such as insufficient integration of affective and cognitive perception, unstable visual perception in complex scenarios, poor temporal consistency of dynamic modeling, and low accuracy of 3D spatial representation. To address these challenges, this report presents a series of research achievements around the core theme of visual affective-cognitive perception and modeling, covering three interrelated and complementary research directions. First, the mechanism of facial emotion analysis and expression is studied, constructing an integrated perception-representation-generation technical system for facial affective-cognitive information. This system includes micro-expression recognition, facial action unit (AU) detection, and emotion transfer, which enables fine-grained affective perception and natural expression, laying a foundation for affective-cognitive interaction between embodied agents and humans. Second, research on feature fusion and temporal modeling for image enhancement and video generation is carried out, establishing a technical path of image enhancement, feature disentanglement, temporal prediction, and video generation. This path realizes visual perception under low-quality conditions and accurate prediction of dynamic scene evolution, providing technical support for the temporal dimension of visual affective-cognitive modeling. Third, the geometric representation and spatiotemporal modeling for object detection and 3D scene reconstruction are explored, building an integrated system of object detection, geometric characterization, 3D reconstruction, and spatiotemporal modeling. This system realizes positioning of complex objects and accurate reconstruction of 3D scenes, constructing a structured spatial representation for visual affective-cognitive modeling. These works lay a foundation for the development of embodied intelligence and the construction of the embodied world model, and provide important technical references for the application of visual affective-cognitive technology in digital humans, service robots, VR/AR and other related fields.

TBD

Keyu Chen

Vivavia Inc.

Abstract：TBD

Common-Individual Embedding for Dynamic Networks with Temporal Group Structure

张妍 [1],Yan Zhang[2],范新妍 [3],方匡南 [2]

1. Shanghai University of International Business and Economics
2. Xiamen University
3. Renmin University of China

Abstract： We propose GANE (Group-Aware Network Embedding), a joint embedding framework for dynamic networks that captures both stable global structures and localized, group-specific temporal variations. To further improve the model's adaptability to transient changes, we introduce Sparse GANE, which models temporal dynamics as sparse perturbations, thereby improving interpretability. Unlike existing methods that either overlook periodic temporal variation or ignore sparsity in dynamic components, GANE leverages temporal clustering to uncover persistent and periodically evolving group structures, while Sparse GANE further enhances interpretability by modeling time-varying structures as sparse perturbations. We also provide non-asymptotic error guarantees of embeddings and show that our estimator can reliably identify changed node pairs when deviations are sparse. The performance of the proposed GANE framework is demonstrated through simulation studies and two real-world applications. Our findings underscore the power of structured dynamic embeddings for revealing interpretable patterns in network evolution.

Latent Space Model under Edge Contamination

Yongqin Qiu[1],Xinyu Zhang[2]

1. University of Science and Technology of China
2. Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Abstract： Network analysis is widely studied across numerous fields, but it typically assumes that observed edges faithfully represent the true underlying relationships. In practice, however, network observations are often contaminated. For instance, in friendship networks, true friends may not appear connected (“false negatives”), while non-friends may be linked for incidental reasons (“false positives”). Consequently, applying standard models directly to such contaminated data leads to inconsistent estimators. To address this issue, we propose a novel latent space model that explicitly accounts for edge contamination by constructing an observed likelihood function. This framework enables parameter estimation without requiring prior knowledge of the noise rates. Computationally, an Expectation–Maximization (EM) algorithm is derived to efficiently maximize the likelihood. Instead of relying on a single optimal tuning parameter, we employ a model averaging method that uses a K-fold cross-validation criterion to automatically assign weights to candidate models. Theoretically, we establish the consistency of the proposed estimator and prove the weight convergence of the model averaging procedure under mild conditions. Simulation studies demonstrate that our method outperforms competing approaches in both predictive accuracy and robustness. We further demonstrate its practical utility through real-world applications in a friendship network and supply chain management.

Likelihood-Based Change Point Detection for Sparse SBM Networks

Yi Ding

School of Statistics, University of International Business and Economics

Abstract: In dynamic networks, connectivity patterns often evolve over time, and usually lead to structural changes in the underlying network structure. Detecting changes, is crucial for understanding network dynamics. Existing methods often struggle in sparse networks, where limited connectivity information reduces detection accuracy. Motivated by this challenge, we propose a likelihood-based framework for detecting change points in dynamic networks that leverages the full structural information of the observed networks. Based on stochastic block models (SBM), we construct a likelihood-based objective function that incorporates both the goodness-of-fit of the observed edges and a penalty on the number of communities within each segment. The proposed framework enables the joint estimation of change point locations and the optimal number of communities for each multi-layer segment. By exploiting the full likelihood of the network structure, the method remains robust even under high sparsity. Extensive numerical experiments, along with applications to the Academic Collaboration Network and the MIT Proximity dataset, demonstrate the effectiveness of the proposed approach in identifying both change points and community structures.

Graphical Models for Mixed-type Data

Yuyang Liu

Shanghai University of International Business and Economics

Abstract: TBD

Statistical ranking with dynamic covariates

Ruijian Han[1],Ruijian Han[1],Binyan Jiang[1],Yiming Xu[2]

1. The Hong Kong Polytechnic University
2. University of Kentucky

Abstract： We introduce a general covariate-assisted statistical ranking model within the Plackett–Luce framework. Unlike previous studies that focus on individual effects with fixed covariates, our model allows covariates to vary across comparisons. This added flexibility enhances model fitting but also brings significant challenges in analysis. This article addresses these challenges in the context of maximum likelihood estimation (MLE). We first provide necessary and sufficient conditions for both model identifiability and the unique existence of the MLE. Then, we develop an efficient alternating maximization algorithm to compute the MLE. Under suitable assumptions on the design of comparison graphs and covariates, we establish a uniform consistency result for the MLE, with convergence rates determined by the asymptotic connectivity of the graph sequence. We also construct random designs under which the proposed assumptions hold almost surely. Numerical studies are conducted to support our findings and demonstrate the model's application to real-world datasets. This is joint work with Pinjun Dong, Binyan Jiang, and Yiming Xu.

Prediction-Oriented Transfer Learning for Survival Analysis

Yu Gu

University of Hong Kong

Abstract： Transfer learning is beneficial for survival analysis, especially when the target study has a limited number of events. However, existing transfer learning methods rely on the restrictive assumption that the target and source studies share similar parameters under Cox models, and most require access to individual-level source data. We propose a novel transfer learning framework that enhances model-based survival prediction by transferring predictive rather than distributional knowledge from source studies. Our approach employs flexible semiparametric transformation models for the target data while eliminating the need to model or share the source data. The ingeniously designed penalty enables simple and stable computation via an EM algorithm. We rigorously establish the asymptotic properties of the proposed estimator and show that it achieves a faster convergence rate than the target-only estimator when source knowledge is sufficiently accurate. We demonstrate the advantages of our methods through extensive simulation studies and an application to two major breast cancer studies.

What Separates Useful from Useless Synthetic Data? Verifying Synthetic Data through Representations

Chendi Wang

Xiamen University

Abstract： Synthetic data is increasingly used in model training and has shown promising performance in supervised fine-tuning and alignment under limited resources. However, naive use of synthetic data can lead to model collapse, in which predictive performance deteriorates substantially. Despite recent empirical mitigation strategies, there remains limited understanding of which synthetic data are beneficial for learning and which are harmful. In this work, we study the quality of synthetic data from a representation-learning perspective. We show that the effectiveness of synthetic data is closely related to the quality of its learned representations, as characterized by Neural Collapse (NC) theory. In particular, synthetic samples exhibiting stronger NC structure can prevent model collapse and ultimately improve generalization. Motivated by this insight, we propose a simple synthetic data filtering method based on pseudo-label confidence thresholding. The proposed approach selectively retains high-confidence synthetic samples and consistently improves model performance. We further provide theoretical results linking this strategy to NC. Experiments on image classification tasks validate our theory.

Homogeneity Pursuit in Ranking Inference Based on Pairwise Comparison

Yuxin Tao[1],Tracy Ke[2]

1. Southern University of Science and Technology
2. Harvard University

Abstract： The Bradley-Terry-Luce (BTL) model is one of the most celebrated models for ranking inference based on pairwise comparison data, ranking individuals by their latent preference scores. A critical question that arises is the uncertainty quantification for ranks. Intuitively, the relative ranks of two individuals become unreliable when their preference scores differ only subtly. In this paper, we explore the homogeneity of preference scores in the BTL model, which assumes that individuals cluster into groups with the same scores. We propose novel penalized maximum likelihood estimators (MLEs) to simultaneously and rigorously perform estimation and clustering. We establish the statistical properties of the proposed methods and develop corresponding inference procedures. By leveraging this group structure, we achieve a faster convergence rate and sharper confidence intervals for the MLE of preference scores, providing new insight into the power of exploiting low-dimensional structures in high-dimensional settings. To address computational challenges, we further develop an ADMM algorithm for efficient optimization. Extensive simulations and real data analyses, including NBA team rankings and statistical journal rankings, demonstrate the improved prediction performance and enhanced interpretability of our model.

Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery

Anderson Ye Zhang

University of Pennsylvania

Abstract: We study maximum likelihood estimation (MLE) in the generalized group orbit recovery model, where each observation is generated by applying a random group action and a known, fixed linear operator to an unknown signal, followed by additive noise. This model is motivated by single-particle cryo-electron microscopy (cryo-EM) and can be viewed primarily as a structured continuous Gaussian mixture model. In practice, signal estimation is often performed by marginalizing over the group using a uniform distribution—an assumption that typically does not hold and renders the MLE misspecified. This raises a fundamental question: how does the misspecified MLE perform? We address this question from several angles. First, we show that in the absence of projection, the misspecified population log-likelihood has desired optimization landscape that leads to correct signal recovery. In contrast, when projections are present, the global optimizers of the misspecified likelihood deviate from the true signal, with the magnitude of the bias depending on the noise level. To address this issue, we propose a joint estimation approach tailored to the cryo-EM setting, which parameterizes the unknown distribution of the group elements and estimates both the signal and distribution parameters simultaneously.

Stable Gaussian Mixture Black-Box Variational Inference

Zhengyu Huang

Peking University

Abstract: Black-box variational inference (BBVI) with Gaussian mixture families offers a flexible approach for approximating complex posterior distributions without requiring gradients of the target density. However, standard numerical optimization methods often suffer from instability and inefficiency. We develop a stable and efficient framework that combines three key components: (1) affine-invariant preconditioning via natural gradient formulations, (2) an exponential integrator that unconditionally preserves the positive definiteness of covariance matrices, and (3) adaptive time stepping to ensure stability and to accommodate distinct warm-up and convergence phases. The proposed approach has natural connections to manifold optimization and mirror descent. For Gaussian posteriors, we prove exponential convergence in the noise-free setting and almost-sure convergence under Monte Carlo estimation, rigorously justifying the necessity of adaptive time stepping. Numerical experiments on multimodal distributions, Neal's multiscale funnel, and a PDE-based Bayesian inverse problem for Darcy flow demonstrate the effectiveness of the proposed method.

Estimation of Out-of-Sample Sharpe Ratio for High Dimensional Portfolio Optimization

Weichen Wang[1],曹原 [2],王炜辰 [2]

1. University of Michigan
2. The University of Hong Kong

Abstract： Portfolio optimization aims at constructing a realistic portfolio with significantout-of-sample performance, which is typically measured by the out-of-sample Sharperatio. However, due to in-sample optimism, it is inappropriate to use the in-sampleestimated covariance to evaluate the out-of-sample Sharpe, especially in the highdimensional settings. In this paper, we propose a novel method to estimate the outof-sample Sharpe ratio using only in-sample data, based on random matrix theory.Furthermore, portfolio managers can use the estimated out-of-sample Sharpe as acriterion to decide the best tuning for constructing their portfolios. Specifically, weconsider the classical framework of Markowitz mean-variance portfolio optimizationunder high dimensional regime of $p/n o c in (0, infty)$, where p is the portfolio dimension and n is the number of samples or time points. We propose to correct thesample covariance by a regularization matrix and provide a consistent estimator ofits Sharpe ratio. The new estimator works well under either of the following conditions: (1) bounded covariance spectrum, (2) arbitrary number of diverging spikeswhen $c < 1$, and (3) fixed number of diverging spikes with weak requirement on theirdiverging speed when $c ge 1$. We can also extend the results to construct global minimum variance portfolio and correct out-of-sample efficient frontier. We demonstratethe effectiveness of our approach through comprehensive simulations and real dataexperiments. Our results highlight the potential of this methodology as a useful toolfor portfolio optimization in high dimensional settings.

Transfer learning in nonparametric online learning problems

Feiyu Jiang

Fudan University

Abstract： This talk focuses on transfer learning for nonparametric online decision-making under covariate shift, with two illustrative applications: contextual bandits and dynamic pricing. In both settings, we demonstrate that leveraging source-domaininformation can significantly reduce regret. We begin with contextual multi-armed bandits in which both target and auxiliary data are protected by local differentialprivacy. We introduce a jump-start estimator and confidence-based policy that effectively integrate heterogeneous source data, together with matching regret upper bounds and minimax lower bounds. We then consider nonparametric dynamic pricing with limited target data. Under Lipschitz reward functions and covariate shift, we propose a transfer-enhanced pricing strategy and establish corresponding regret bounds and lower bounds. Taken together, these studies quantify when andhow rich source-domain samples can be harnessed to improve statistical efficiencyin online learning.

Slacked Empirical Likelihoods for Post-Criterion Inference

Yiyuan She

Westlake University

Abstract： Statistical inference under nonsmooth penalties presents significant challenges for traditional empirical likelihood methods developed in low dimensions. We introduce SEL from the penalized criterion itself and use slack variables to convert its structural conditions into a tractable dual formulation over a family of divergence functions. This makes penalty-induced structural bias explicit, yielding a transparent distinction between noncentral and central limiting behavior. In particular, SEL introduces a data-driven dual centering scheme that cancels the bias terms and recovers a central chi-square limit. The resulting theory covers classical and high-dimensional regimes, attains sample-complexity scalings standard in sparse inference.

Bayesian reinforcement learning framework for optimizing the BCI-utility of P300 Brain-Computer Interfaces

Bangyao Zhao,Jane E. Huggins,Yixin Wang,Jian Kang

University of Michigan

Abstract： Brain-computer interfaces (BCIs) enable direct communication between the brain and computers, providing critical tools for people with disabilities to communicate with the world. The performance of BCIs is often evaluated using BCI-utility, a comprehensive metric that balances both accuracy and speed in communication. This paper introduces a Bayesian reinforcement learning framework to optimize the BCI-utility of the P300 BCI, a BCI system that identifies a user's intended character on a virtual keyboard by analyzing EEG responses to stimuli. We construct confidence scores for each character based on EEG responses and then propose a unified learning framework that explicitly maximizes BCI-utility. It integrates two key components: an early stopping policy and a dynamic stimulus selection policy. The early stopping policy is optimized using an actor-critic algorithm, while a Gaussian process-based Bayesian model is developed to learn transition dynamics to guide the selection of the next stimulus. The proposed framework effectively addresses critical implementation challenges, including pauses between characters, double-target issues, and delays caused by the time required for EEG responses. Extensive simulations under varying signal-to-noise ratios (SNRs) and evaluations on recorded human EEG data demonstrate that our method significantly improves BCI-utility compared to existing approaches. This work highlights the potential of reinforcement learning to improve the performance and usability of P300 BCI systems.

Online Sequential Decision-Making with Reinforcement Learning: From Robotics to Human-centered tasks

Ran Chen

Washington University in St. Louis

Abstract: Many decision-making tasks in business and healthcare—such as dynamic assortment and pricing in operations management, or selecting treatments for patients over time in clinical care—can be naturally framed as online sequential decision-making problems. Reinforcement learning (RL), originally developed for engineering applications, provides a powerful framework for tackling these challenges. However, human-centric tasks introduce new challenges, including continuous and heterogeneous data, high-dimensional data, the need for interpretability, and demands for personalization in fully online settings.

In this talk, I will present my work on developing personalized RL algorithms for sepsis treatment in the ICU, demonstrating how RL can support real-time, patient-specific medical decisions. Time permitting, I will also briefly introduce my work on addressing high-dimensionality in business decision-making, as well as related work in mobile health applications.

Inference on Large-scale Partially Functional Linear Model with Heterogeneous Errors

Kaijie Xue

Shanghai University of International Business and Economics

Abstract: We investigate a partially functional linear model by focusing on the heterogenous error scenario in which the scalar response is associated with an ultra-large number of both functional predictors and scalar covariates. Moreover, the model does not require the standard condition on eigenvalue decay for functional predictors, leading to a more challenging and general framework. The target is to establish a rigorous inferential procedure for hypothesis testing on an arbitrary subset of both regression functions and scalar coefficients. Specifically, we devise a confidence region for post-regularization inference using a pseudo score function that is not decorrelated owing to the heterogenous errors. The proposed test does not require estimation consistency of the functional part, and is shown to be uniformly convergent to the prescribed significance. We investigate the finite-sample performance of the proposed model using simulation studies and an application to functional magnetic resonance imaging brain image data.

Identification and Inference for Structural Accelerated Failure Time Models via Instrument Interactions

Xingqiu Zhao

The Hong Kong Polytechnic University

Abstract： We study causal inference for right-censored time-to-event outcomes under unmeasured confounding using structural accelerated failure time models. Leveraging interactions among instrumental variables, our framework enables identification and inference without requiring classical instrument validity. We address right censoring via an augmented inverse probability censoring weighting approach, yielding a Neyman-orthogonal moment function with double robustness. Estimation uses generalized empirical likelihood, suited for many potentially weak moment conditions. We establish consistency and asymptotic normality under many-weak-moment asymptotics and provide diagnostic tools. Simulations show strong finite-sample performance across various censoring rates and instrument configurations. An application to UK Biobank data demonstrates practical utility for large-scale observational survival analysis.

Neural frailty machines for survival analysis

Wen Yu

Fudan University

Abstract： 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.

Variable Significance Testing for the Deep Cox Model

Qixian Zhong

Xiamen University

Abstract: Deep learning has become enormously popular in the analysis of complex data, including event time measurements with censoring. To date, deep survival methods have mainly focused on prediction. Such methods are scarcely used in matters of statistical inference such as hypothesis testing. Due to their black-box nature, deep-learned outcomes lack interpretability which limits their use for decision-making in biomedical applications. This paper provides estimation and inference methods for the nonparametric Cox model -- a flexible family of models with a nonparametric link function to avoid model misspecification. Here we assume the nonparametric link function is modeled via a deep neural network. To perform statistical inference, we utilize sample splitting and cross-fitting procedures to get neural network estimators and construct test statistic. These procedures enable us to propose a new significance test to examine the association of certain covariates with event times. We establish convergence rates of the neural network estimators, and show that deep learning can overcome the curse of dimensionality in nonparametric regression by learning to exploit low-dimensional structures underlying the data. In addition, we show that our test statistic converges to a normal distribution under the null hypothesis and establish its consistency, in terms of the Type II error, under the alternative hypothesis. Numerical simulations and a real data application demonstrate the usefulness of the proposed test.

Deep partially linear transformation model for right-censored survival data

Zhangsheng Yu

Shanghai Jiao Tong University

Abstract: Although the Cox proportional hazards (PH) model is well established and extensively used in the analysis of survival data, the PH assumption may not always hold in practical scenarios. The class of semiparametric transformation models extends the Cox model and also includes many other survival models as special cases. This paper introduces a deep partially linear transformation model as a general and flexible regression framework for right-censored data. The proposed method is capable of avoiding the curse of dimensionality while still retaining the interpretability of some covariates of interest. We derive the overall convergence rate of the maximum likelihood estimators, the minimax lower bound of the nonparametric deep neural network estimator, and the asymptotic normality and the semiparametric efficiency of the parametric estimator. Comprehensive simulation studies demonstrate the impressive performance of the proposed estimation procedure in terms of both the estimation accuracy and the predictive power, which is further validated by an application to a real-world dataset.

TBD

Yongfu Yu

Fudan University

Abstract：TBD

Doubly Robust Estimators for Heterogeneous Treatment Effects in Heteroskedastic Survival Data and Application

Fangyao Chen

Xi'an Jiaotong University

Abstract： With the advancement of precision medicine, numerous statistical approaches have been developed to estimate heterogeneous treatment effects (HTEs). While medical research commonly encounters survival outcomes subject to inherent censoring and heteroscedasticity, for which existing HTE estimation methods exhibit notable limitations. Neglecting censoring and heteroscedasticity may introduce substantial bias. To address this issue, we propose two doubly robust (DR) methods for HTE estimation based on nonparametric failure time (NFT) Bayesian additive regression trees (BART). First, we employ NFT-BART as the core prediction model to relax restrictive assumptions, including linearity, proportional hazards and homoscedasticity. Then, we extend the DR-Learner framework to survival settings, enabling effective handling of censoring and confounding inherent in observational data. Finally, we design a series of data-generating processes to conduct comprehensive simulation experiments. We systematically evaluate how HTE estimation is affected by multiple key factors including training sample size, propensity score distribution, censoring rate, imbalanced treatment allocation, complexity of model and bias functions, as well as homoscedastic and heteroscedastic outcome structures. Simulation results verify that our two proposed methods yield satisfactory efficiency and robustness for HTE estimation. Moreover, utilizing real-world observational data from the National Health and Nutrition Examination Survey (NHANES), we demonstrate the practical utility of our methods in personalized hypertension management. In summary, the newly developed approaches enable robust, accurate and reliable HTE quantification for observational survival data.

Distributed Censored Quantile Regression: Convolution Smoothing and Communication Efficiency

Huijuan Ma[2],焦小奇 [1],马慧娟 [1],张宝学 [1]

1. East China Normal University
2. Capital University of Economics and Business

Abstract： Censored quantile regression (CQR) has become a popular framework for analyzing survival outcomes, yet conventional estimation procedures face computational challenge when scaled to massive datasets. In such modern large-scale settings, individual-level survival data cannot be freely accessed or centrally pooled due to privacy concerns and storage limitations, rendering distributed computation increasingly indispensable. The non-differentiability of the CQR loss further complicates optimization, making existing methods computationally expensive and poorly suited for distributed data structures. To overcome these challenges, we propose a communication-efficient distributed CQR framework that leverages convolution smoothing to construct a globally smooth and convex objective, thereby enabling quasi-Newton optimization in large-scale distributed environments. The proposed approach effectively balances computational scalability and communication efficiency. We establish the Bahadur representation and asymptotic properties of the convolution-smoothed CQR estimator on a single machine, and further derive the convergence rate of the distributed quasi-Newton updates. Extensive simulations and real-world applications demonstrate that the proposed method achieves statistical accuracy comparable to single-machine full-sample algorithms, while outperforming existing distributed CQR methods in both computational and communication efficiency.

Dynamic-Centime：一种利用纵向数据预测剩余生存时间的动态预测模型

Chengfeng Zhang[1],陈征 [1,2]

1. Southern Medical University
2. 多器官损伤防治全国重点实验室

Abstract：准确预测患者的剩余生存时间对于个性化医疗至关重要。然而，当前的电子健康档案（EHR）数据普遍存在非独立删失和高删失率的问题，给传统的生存分析方法带来了巨大挑战。为解决这一问题，我们提出了一种新型动态生存预测模型——Dynamic-Centime。该方法通过引入循环神经网络处理纵向协变量，并将基础的 Centime 模型拓展为条件 Centime 架构，从而能够持续整合新观测数据，在不同时间点动态预测患者的剩余生存时间。模拟研究表明，在不同的删失比例、协变量交互效应及非比例风险等复杂场景下，Dynamic-Centime 的预测性能和 TransformerJM 性能接近，并且优于 Match 方法。此外，我们将该模型应用于真实的重症脓毒症患者队列，结果显示，在个体预测层面，Dynamic-Centime 不仅可以估计预期剩余生存时间，还可动态输出未来 w 天的条件生存概率，其预测性能接近 TransformerJM 且优于 Match 方法。总之，本文提出的 Dynamic-Centime 模型有效解决了 EHR 中纵向生存数据的复杂性问题，为临床医生在复杂病程中做出精准医疗决策提供了坚实的科学支持。

Self-Evolving Agents: Boosting Efficiency and Capability via Architectural Progression

Chi Zhang
Westlake University

Abstract：TBD

Towards Efficient AI: Optimizing Deployment and Inference

Wangbo Zhao

The Hong Kong University of Science and Technology

Abstract： As modern AI models continue to scale, their capabilities have improved dramatically; however, the exponential growth in model size has significantly outpaced linear hardware advancement, creating an "efficiency gap". This gap poses critical challenges for broad AI adoption, such as environmental sustainability, resource constraints on edge devices, and strict latency requirements for real-time applications. To bridge this gap, this presentation focuses on Efficient AI, systematically optimizing the AI lifecycle across two pivotal stages: Deployment and Inference.

This talk will delve into two core research dimensions:

- Accessible & Flexible Deployment: We address the challenge of massive model sizes through advanced structural pruning and elastic architectures. Key contributions include MoNE, which replaces redundant experts in MoE models with lightweight novices, and EA-ViT, which enables vision transformers to adapt their active size and complexity to diverse target devices.
- Dynamic & Rapid Inference: To alleviate latency bottlenecks, we explore dynamic computation to eliminate redundancy. We introduce the DyDiT series, which implements timestep and spatial-wise dynamics for diffusion transformers. Furthermore, we present RAPID3, a reinforced acceleration policy for training-free efficiency, and SGL, which uses small VLMs to guide large model inference.

Finally, I outline a future research vision centered on Model-System Co-Design. By deepening the interaction between model architectures and system-level optimizations, we aim to build the next generation of truly sustainable and democratized Efficient Intelligence.

Analogy, Abstraction, and Reasoning in Multimodal Large Models

Xu Yang

Southeast University

Abstract: In recent years, large multimodal models (LMMs) have made significant progress by building on the strengths of large language models (LLMs). As LLMs have evolved, they have demonstrated strong emergent abilities in analogy, abstraction, and reasoning, and LMMs have inherited many of these capabilities. However, because LMMs rely heavily on the reasoning abilities of LLMs, a series of issues have emerged in multimodal analogy and reasoning—such as shortcut reasoning and incompatibilities between visual and language modules.

This talk focuses on several problems identified in recent research on multimodal analogy, particularly in the context of in-context learning, and discusses preliminary approaches for addressing these challenges. For shortcut reasoning, the talk presents an in-context-vector-based method that extracts common task patterns from a small number of examples to mitigate the tendency toward shortcuts. For the issue of visual–textual incompatibility, the talk introduces a simple strategy that first enhances textual reasoning ability and then strengthens visual reasoning capability.

Towards Efficient Inference of Large Foundation Models

Bohan Zhuang

Zhejiang University

Abstract: This talk will focus on cutting-edge efficient inference techniques for foundation models, covering multimodal LLMs and video world models. It will also share our team's research advances in algorithm-system co-design, along with key insights into the future evolutionary trajectory of efficient foundation models.

Hierarchical Contrastive Learning for Multimodal Data with Partial Sharing

Doudou Zhou

National University of Singapore

Abstract: Most multimodal representation learning methods rely on a shared-versus-private decomposition, treating latent information as either common to all modalities or specific to one. This view is often too coarse: in many applications, important factors are shared by only a subset of modalities, creating partially shared structure that standard methods do not model explicitly. Ignoring this intermediate layer can lead to over-alignment of unrelated signals, loss of complementary information, and weaker downstream performance. To address this problem, we propose Hierarchical Contrastive Learning (HCL), a framework for learning multimodal representations with globally shared, partially shared, and modality-specific components. HCL is built on a hierarchical latent-variable model with structural sparsity and a structure-aware contrastive objective that aligns only the modalities that truly share a latent component. Under independence of the latent variables, we establish identifiability of the hierarchical decomposition up to orthogonal transformation, prove recovery guarantees for the loading matrices under gradient descent, and analyze downstream prediction through debiased and group-regularized linear regression with estimation and excess-risk bounds. Simulations demonstrate accurate recovery of hierarchical latent structure and effective identification of task-relevant components, while real-data analysis on multimodal electronic health records shows that explicitly modeling partial sharing leads to more informative representations and improved predictive performance.

Semi-supervised Clustering Through Representation Learning of High-dimensional Count Data

Mengyan Li

Bentley University

Abstract: High-dimensional count data arise in many modern applications, where observations are sparse, heterogeneous, and driven by complex latent structure. These characteristics make modeling and prediction challenging, especially when labels are limited or partially observed. We propose SCORE, a semi-supervised representation learning framework for clustering and embedding construction from high-dimensional count features. SCORE is built on a Poisson-adapted latent factor mixture model that supports incorporating external pre-trained feature embeddings when available, enabling efficient characterization of feature patterns and extraction of meaningful latent cluster membership and low-dimensional representations. To scale inference and learning to large datasets, SCORE uses a hybrid algorithm that combines expectation maximization with Gaussian variational approximation, leveraging a small labeled subset to refine estimation on a large pool of unlabeled samples. We establish convergence guarantees for this hybrid procedure, quantify approximation errors from Gaussian variational approximation, and derive error rates under increasing embedding dimensions. Our theory and experiments show that incorporating unlabeled data improves accuracy and reduces sensitivity to label scarcity, yielding strong finite sample performance relative to existing approaches. We demonstrate SCORE on electronic health record count features for predicting disability status in multiple sclerosis, where it learns informative and predictive patient embeddings, illustrating its practical value.

Multi-view Spherical Mixture Models for Aligning Partially Overlapping Feature Spaces with Latent Synonymy

Yuming Zhang[1],Congyuan Duan[2],Dong Xia[2],Doudou Zhou[3],Tianxi Cai[1]

1. Harvard University
2. Hong Kong University of Science and Technology
3. National University of Singapore

Abstract: Multi-institutional electronic health record (Multi-EHR) data have emerged as a powerful resource for developing predictive models to support clinical decisions and for generating reliable real-world evidence. By aggregating information from diverse patient populations and institutions, they enhance the robustness and generalizability of models and findings. However, analyzing multi-EHR remains challenging because disparate institutions rarely map all data elements to common ontology, and raw EHR codes are often overly granular and institution-specific, fragmenting representations of the same clinical concept. Hence, integrative analysis must overcome two key hurdles: harmonizing codes with the same clinical meaning (synonymy), and aligning institutional feature spaces. To address these challenges, we propose a multi-view spherical mixture framework for joint alignment and synonym recovery, where embeddings from heterogeneous sources serve as privacy-preserving summaries of clinical concepts. Synonymy is modeled via a mixture of von Mises-Fisher distributions, yielding unified representations that consolidate semantically equivalent raw codes. We develop a composite likelihood estimation procedure and establish non-asymptotic error bounds for latent representations and mixture mean directions, together with consistent recovery of synonym clusters. The theory quantifies statistical gains from integrating multiple sources and auxiliary knowledge graph information. Simulations and a multi-institutional EHR application demonstrate improved alignment and synonym clustering.

Efficient inference on high-dimensional logistic regression under class imbalance

Alexander Giessing

National University of Singapore

Abstract: Standard debiasing methods for high-dimensional logistic regression perform poorly under severe class imbalance, which is typical in rare-event prediction, from disease screening to credit default. We propose the Gauss--Markov debiasing program, a framework for constructing approximately unbiased, minimum-variance one-step estimators that remain valid and efficient under imbalance. The method supports inference for conditional log-odds and case probabilities at user-specified query points without sparsity assumptions on the query point, attains the semiparametric efficiency bound asymptotically, and reduces finite-sample bias and variance inflation. We illustrate the method by building a Type II diabetes risk model using demographic and genome-wide genetic data from the Mass General Brigham Biobank.
This is joint work with Wenjie Guan (Cornell University), Yikun Zhang (University of Washington), Doudou Zhou (National University of Singapore), and Tianxi Cai (Harvard University).

Deep Learning Assisted Variable Selection with False Discovery Rate Control

Changcheng Li

Dalian University of Technology

Abstract: Variable selection in high-dimensional nonlinear settings remains challenging, especially when controlling the false discovery rate (FDR) without imposing restrictive model assumptions. We propose a model-free variable selection framework that integrates deep learning and achieves asymptotically valid FDR control. Our approach uses gradients of a trained neural network with respect to the input variables as importance measures, and employs sample splitting to construct artificial null variables for calibration. We show that the proposed procedure asymptotically controls the FDR under appropriate conditions. Empirical studies on synthetic datasets demonstrate that the method effectively identifies true signals while maintaining reliable FDR control, even in the presence of nonlinearity and feature correlation.

Maximin Learning of Individualized Treatment Effect on Multi-Domain Outcomes

Molei Liu

Peking University

Abstract: Precision mental health requires treatment decisions that account for heterogeneous symptoms reflecting multiple clinical domains. However, existing methods for estimating individualized treatment effects (ITE) rely on a single summary outcome or a specific set of observed symptoms or measures, which are sensitive to symptom selection and limit generalizability to unmeasured yet clinically relevant domains. We propose DRIFT, a new maximin framework for estimating robust ITEs from high-dimensional item-level data by leveraging latent factor representations and adversarial learning. DRIFT learns latent constructs via generalized factor analysis, then constructs an anchored on-target uncertainty set that extrapolates beyond the observed measures to approximate the broader hyper-population of potential outcomes. By optimizing worst-case performance over this uncertainty set, DRIFT yields ITEs that are robust to underrepresented or unmeasured domains. We further show that DRIFT is invariant to admissible reparameterizations of the latent factors and admits a closed-form maximin solution, with theoretical guarantees for identification and convergence. In analyses of a randomized controlled trial for major depressive disorder (EMBARC), DRIFT demonstrates superior performance and improved generalizability to external multi-domain outcomes, including side effects and self-reported symptoms not used during training.

Double Fairness Policy Learning: Integrating Action Fairness and Outcome Fairness in Decision-making

Zeyu Bian

FSU

Abstract: Fairness is a central pillar of trustworthy machine learning, especially in domains where accuracy- or profit-driven optimization is insufficient. While most fairness research focuses on supervised learning, fairness in policy learning remains less explored. Because policy learning is interventional, it induces two distinct fairness targets: action fairness (equitable action assignments) and outcome fairness (equitable downstream consequences). Crucially, equalizing actions does not generally equalize outcomes when groups face different constraints or respond differently to the same action.

We propose a novel double fairness learning (DFL) framework that explicitly manages the trade-off among three objectives: action fairness, outcome fairness, and value maximization. We integrate fairness directly into a multi-objective optimization problem for policy learning and employ a lexicographic weighted Tchebyshev method that recovers Pareto solutions beyond convex settings, with theoretical guarantees on the regret bounds. Our framework is flexible and accommodates various commonly used fairness notions. Extensive simulations demonstrate improved performance relative to competing methods. In applications to a motor third-party liability insurance dataset and an entrepreneurship training dataset, DFL substantially improves both action and outcome fairness while incurring only a modest reduction in overall value.

Identifying the Desert Decision Rule to Assess and Achieve Fairness

Ping Zhang

Peking University

Abstract: We study fairness in decision-making when the data may encode systematic bias. Existing approaches typically impose fairness constraints while predicting the observed decision, which may itself be unfair. We propose a novel framework for characterising and addressing fairness issues by introducing the notion of desert decision, a latent variable representing the decision an individual rightfully deserves based on their actions, efforts, or abilities. This formulation shifts the prediction target from the potentially biased observed decision to the desert decision. We advocate achieving fair decision-making by predicting the desert decision and assessing unfairness by the discrepancy between desert and observed decisions. We establish nonparametric identification results under causally interpretable assumptions on the fairness of the desert decision and the unfairness mechanism of the observed decision. For estimation, we develop a sieve maximum likelihood estimator for the desert decision rule and an influence-function-based estimator for the degree of unfairness. Sensitivity analysis procedures are further proposed to assess robustness to violations of identifying assumptions. Our framework connects fairness with measurement error models, aligning predictive accuracy with fairness relative to an appropriate target, and providing a structural approach to modelling the unfairness mechanism.

Locally Private Estimation with Public Features

Yuheng Ma,贾珂,杨翰方

Renmin University of China

Abstract： We initiate the study of locally differentially private (LDP) learning with public features. We define semi-feature LDP, where some features are publicly available while the remaining ones, along with the label, require protection under local differential privacy. Under semifeature LDP, we consider three fundamental estimation problems: non-parametric density estimation, classification, and regression. Given the smoothness assumption, we show that the minimax convergence rate is significantly improved compared to classical LDP. Then, we propose HistOfTree, an estimator that fully leverages the information contained in both public and private features. Theoretically, HistOfTree reaches the mini-max optimal convergence rate. Empirically, HistOfTree achieves superior performance on both synthetic and real data. We also explore scenarios where users have the flexibility to select features for protection manually. In such cases, we propose an estimator and a data-driven parameter tuning strategy, leading to analogous theoretical and empirical results.

Second-Order Sparse Sufficient Dimension Reduction with Applications to Quadratic Discriminant Analysis

Jing Zeng

University of Science and Technology of China

Abstract： Motivated by exploratory data analysis, sufficient dimension reduction (SDR) methods, especially inverse regression methods such as sliced inverse regression (SIR) and sliced averaged variance estimation (SAVE), have been central to multivariate analysis for more than three decades. Despite their popularity, the extension of these methods to high-dimensional settings remains challenging. This paper addresses the computational and theoretical limitations of the less explored second-order SDR methods in high dimensions. We introduce a novel approach for sparse subspace estimation that utilizes quadratic convex optimization and leverages the group structure of tensor parameters, achieving significant parameter reduction. The proposed two-step estimator achieves consistency in dimension selection, variable selection, and subspace estimation at a high convergence rate under mild conditions. The effectiveness and efficiency of the proposed method are further demonstrated through extensive simulation studies and real data examples. Additionally, the proposed sparse second-order SDR techniques are applied to quadratic discriminant analysis (QDA) problems and provide practitioners a sparse projective classification method that is theoretically guaranteed and empirically well-performed.

Metric conformal prediction based on the expected local radius

Rui Qiu

Peking University

Abstract: We develop a novel metric conformal prediction method to address the challenges of uncertainty quantification for complex non-Euclidean data. Central to our approach is the introduction of the expected local radius, a geometrically interpretable quantity that characterizes the "metric effort'' required to accumulate probability mass around a candidate point. This quantity measures the local concentration of the distribution in abstract metric spaces, thereby enabling the conformal score to produce adaptive prediction sets. Furthermore, we propose a metric distributional random forest to estimate the conditional distribution adaptively, effectively alleviating the curse of dimensionality associated with classical kernel smoothing methods. Theoretically, we establish the uniform consistency of our estimators and prove the asymptotic conditional validity of the resulting prediction sets. Empirical results demonstrate that our method achieves superior performance in scenarios involving multidimensional covariates and complex response structures.

Combining pre-trained large models via localized model averagin

Ziwen Gao

Tsinghua University

Abstract: Many pre-trained large models (PLMs) are being released. A popular approach to taking advantage of such PLMs is fine-tuning. However, when the data suitable for the task of interest are rather limited, fine-tuning may not be effective. In this case, weighting the predictions from different PLMs can be a better way to improve predictive performance. Motivated by such applications, we propose a localized model averaging method with weights modeled as functions of the covariates, making it substantially more versatile than existing model averaging methods. This formulation allows the model averaging procedure to adaptively capture the varying relative advantages of different PLMs across heterogeneous contexts. Specifically, we learn flexible local weights under a general loss framework that accommodates a broad class of prediction tasks. We further establish the asymptotic optimality of the proposed method for both in-sample and out-of-sample risks, as well as the consistency of the estimated weights. Extensive numerical experiments further demonstrate the effectiveness of the proposed method.

Heritability Estimation via Genetic Similarity Representation: Theory and Biobank-Scale Computation

Jianqiao Wang

Tsinghua University

Abstract: Heritability estimation in genome-wide association studies (GWAS) remains challenging in the presence of high dimensionality, heterogeneous genetic effects, and linkage disequilibrium. We introduce SMILE, a similarity-representation framework for robust heritability estimation that models outcome similarity through a weighted genetic similarity matrix. The proposed framework relaxes restrictive assumptions commonly imposed by fixed- and random-effects methods, includes the classical random-effects model as a special case, and avoids the need to accurately estimate regression coefficients or precision matrices. We also discuss biobank-scale implementation of the method, including new algorithms tailored to large-scale computation and GPU architectures. Together, these developments provide a statistically robust and computationally scalable approach to heritability estimation for modern genetic studies.

Generalizing Experience for Language Agents with Hierarchical MetaFlows

Xin Cong

Tsinghua University

Abstract: Recent efforts to employ large language models (LLMs) as agents have demonstrated promising results in a wide range of multi-step agent tasks. However, existing agents lack an effective experience reuse approach to leverage historical completed tasks. In this paper, we propose a novel experience reuse framework MetaFlowLLM, which constructs a hierarchical experience tree from historically completed tasks. Each node in this experience tree is presented as a MetaFlow which contains static execution workflow and subtask required by agents to complete dynamically. Then, we propose a Hierarchical MetaFlow Merging algorithm to construct the hierarchical experience tree. When accomplishing a new task, MetaFlowLLM can first retrieve the most relevant MetaFlow node from the experience tree and then execute it accordingly. To effectively generate valid MetaFlows from historical data, we further propose a reinforcement learning pipeline to train the MetaFlowGen. Extensive experimental results on AppWorld and WorkBench demonstrate that integrating with MetaFlowLLM, existing agents (e.g., ReAct, Reflexion) can gain substantial performance improvement with reducing execution costs. Notably, MetaFlowLLM achieves an average success rate improvement of 32.3% on AppWorld and 6.2% on WorkBench, respectively.

Conformal Inference for Minority Subgroups via Cross-Group Borrowing

Huaqing Jin

Tsinghua University

Abstract： Conformal prediction offers a distribution-free framework for constructing prediction sets with guaranteed coverage. However, existing conformal methods often fail to provide reliable inference for underrepresented minority subgroups due to potential heterogeneity across subpopulations. We propose a new framework, termed Subgroup Conformal Prediction (SCP), to improve prediction performance for minority subgroups by borrowing information from other subgroups. This is achieved by framing subgroup heterogeneity as a general dataset shift problem and applying reweighting strategies to approximate the distribution in the minority subgroup. SCP can be implemented using either a semiparametric exponential tilt model or machine learning classifiers to obtain the weights. We establish theoretical coverage guarantees for both approaches under mild conditions, and show that SCP remains valid even when the working model for the nonconformity score is misspecified. Extensive simulations demonstrate that SCP yields valid and efficient prediction sets, even under substantial sample size imbalance between subgroups. We further evaluate SCP using the Communities and Crimes dataset, focusing on crime rate prediction for African American-majority communities. The results demonstrate that SCP substantially improves prediction performance across multiple crime types, whereas existing conformal methods often fail to achieve nominal coverage.

Design-based inference for edge-level outcomes in directed networks

Hanzhong Liu,Haoyang Yu,Xin Lu

Tsinghua University

Abstract： We study design-based causal inference for edge-level outcomes in directed networks under dyadic interference. In this setting, outcomes are defined on directed edges and depend on the joint treatment assignments of pairs of units, inducing a dependence structure that invalidates standard estimation and inference methods developed for node-level data. Horvitz--Thompson estimators are constructed for a general class of causal effects, and their asymptotic normality is established under mild regularity conditions. To enable valid inference, variance estimators are constructed by exploiting identifiable components of network dependence, yielding substantially less conservative bounds than classical approaches. To improve efficiency, auxiliary covariates are incorporated through a prediction-powered inference framework. A key technical challenge is that standard two-fold sample splitting fails in the presence of dyadic outcomes. We address this by introducing a three-fold sample splitting scheme that restores the conditional independence required for unbiased estimation. Under a stability condition, the resulting adjusted estimator is asymptotically normal and accommodates both linear adjustment and flexible machine learning methods. A calibration step is further introduced to guarantee no asymptotic efficiency loss relative to the unadjusted estimator. Simulation results confirm the theoretical findings and demonstrate substantial efficiency gains.

Towards Intelligent Story Visualization and Editing

Xiaodong Cun

Great Bay University

Abstract: Storytelling has long been a central goal of visual media, and recent advances in diffusion models and multimodal large language models are pushing story visualization beyond single-shot, single-prompt generation toward long, multi-scene, editable narratives. This talk connects three of our recent works along this trajectory. DiTCtrl (CVPR 2025) investigates the attention behavior of the Multi-Modal Diffusion Transformer and introduces a tuning-free attention-control scheme that enables coherent multi-prompt longer video generation, providing a backbone for narrative-driven synthesis. FairyGen (SIGGRAPH Asia 2025) turns a single child-drawn character into a story-driven cartoon video by decoupling character modeling from background generation, and by combining MLLM-based story planning with style-consistent scene construction and style-propagated motion, so that the original artistic style is faithfully preserved. CutClaw (arXiv 2026) shifts from generation to editing, proposing an autonomous multi-agent framework that turns hours-long raw footage and background music into rhythm-aligned short videos through structured captioning, shot planning, and music synchronization. The talk concludes with open challenges in long-range consistency, controllable stylization, and human–agent co-creation on the path from single-shot generation toward narrative-level authoring.

Intelligent Design Generation: From SVG to Parametric CAD

Qian Yu

Beihang University

Abstract: As the digital design paradigm shifts toward intelligence, there is a growing demand for structured design assets—such as SVG vector graphics and parametric CAD models—that support precise editing and lossless scaling. However, conventional creation of such content relies heavily on manual effort and specialized tools, suffering from high entry barriers, low efficiency, and limited automation. To address these challenges, this talk presents our team's progress in intelligent design generation. Leveraging the inherent "visual–code" bimodal nature of structured design data, we propose a series of novel methods that enable efficient and controllable generation spanning 2D/3D/animated SVGs to complex parametric CAD models, significantly advancing the automation, quality, and editing flexibility of structured design asset creation.

Research on Data and Methods for Ultra-High-Definition Visual Generation

Ying Tai

Nanjing University

Abstract： With the widespread adoption of 4K/8K devices, the demand for ultra-high-resolution (UHR) content in film, advertising, and digital economy scenarios is rapidly increasing. This report systematically presents the recent work of our research group on ultra-high-definition visual generation from four perspectives: UHR data, underlying architectures, model design, and acceleration methods.

Specifically, regarding data, the report introduces the construction process of high-quality visual datasets for image generation, image editing, and video generation. In terms of underlying architectures and model design, for images, the report presents a pixel diffusion architecture, DiP, which incorporates local information modeling, and a text-to-image method L2P based on DiP; for videos, it introduces an ultra-high-definition video generation model, LUVE. Finally, regarding acceleration, the report presents a KV Cache–based acceleration method for autoregressive diffusion models, which effectively mitigates error accumulation in autoregressive architectures and demonstrates its effectiveness across text-to-video, image-to-video, and world model tasks. The codes for the related methods have been open-sourced at: https://github.com/NJU-PCALab/.

Multi-View 3D Reconstruction with Radiance Fields

Chunxia Xiao

Wuhan University

Abstract： Multi-View 3D Reconstruction with Radiance Fields is an effective 3D modeling and representation method, and advances the progress of the fields requiring high-precision modeling and realistic rendering, such as autonomous driving simulation, digital asset creation, and film production. This talk first discusses the main challenges the radiance fields facing for multi-view 3D reconstruction, then introduces several methods presented to address these challenges, finally, gives the potential future research directions on the topic.

模拟学习方法论：理论、算法及应用

Jun Shu

Xi'an Jiaotong University

Abstract：近几年，人工智能研究的突破之一是以 ChatGPT/DeepSeek 为代表的大模型的显著发展。相比于以解决特定任务为特征的传统深度学习模型，大模型在解决跨任务泛化等复杂任务中展示出惊人的能力（一般认为是由大模型智能涌现所引起的）。但大模型 "蛮力出奇迹"的实现模式与"资源稀疏型"的学术研究模式存在鸿沟。本报告将从大模型的约化这一问题展开讨论，介绍基于"模型学习方法论"（SLeM）的元学习框架，阐述背后关于任务迁移泛化的统计学习理论，并以机器学习自动化作为典型应用场景展示，研发了一系列机器学习自动化基础算法簇，揭示 SLeM 学习范式对现实场景中的潜在适用性。

Strongly consistent community detection in popularity adjusted block models

袁泉 [2],刘秉辉 [1],Danning Li[1],Lingzhou Xue[3]

1. Northeast Normal University
2. Yunnan University
3. Penn State University

Abstract：The Popularity Adjusted Block Model (PABM) provides a flexible framework for community detection in network data by allowing heterogeneous node popularity across communities. However, this flexibility increases model complexity and raises key unresolved challenges, particularly in effectively adapting spectral clustering techniques and efficiently achieving strong consistency in label recovery. To address these challenges, we first propose the Thresholded Cosine Spectral Clustering (TCSC) algorithm and establish its weak consistency under the PABM. We then introduce the one-step Refined TCSC algorithm and prove that it achieves strong consistency under the PABM, correctly recovering all community labels with high probability. We further show that the two-step Refined TCSC markedly accelerates clustering error convergence, especially with small sample sizes. Additionally, we propose a data-driven approach for selecting the number of communities, which outperforms existing methods under the PABM. The effectiveness and robustness of our methods are validated through extensive simulations and real-world applications.

Optimal Mixture-of-Experts Model Averaging for Conditional Generative Learning

Baihua He

The University of Science and Technology of China

Abstract： Learning complex conditioned distributions is essential in modern generative modeling, powering applications from probabilistic prediction to AI-driven image and text synthesis. Different generative models often excel on some conditioning inputs but struggle on others, yet classical ensemble methods assign fixed weights that ignore this variability. We introduce a statistically principled Optimal Mixture-of-Experts framework that learns input-dependent weights to combine multiple conditional generators dynamically. By extending model averaging theory and using integral probability metrics to align distributions, we prove our adaptive weighting achieves asymptotic optimality under broad conditions. The proposed method can seamlessly incorporate any conditional generator, and can be extended to other probabilistic tasks. Comprehensive simulations and real-world experiments demonstrate that our method consistently outperforms individual models and traditional ensembles. These experiments, such as image generation and demand distribution learning, show its broad effectiveness in other statisitic and AI applications.

Controlling the False Discovery Rate in High-Dimensional Linear Models Using Model-X Knockoffs and p-values

常晋源 [1],李晨龙 [1],汤琤咏 [2],朱正天 [3]

1. Southwestern University of Finance and Economics
2. Temple University
3. Tongji University

Abstract： We propose a novel multiple testing methodology for controlling the false discovery rate (FDR) in high-dimensional linear models that integrates model-X knockoff techniques with debiased penalized regression estimators. At the foundation of our methodology, we construct and study two sets of naturally paired high dimensional test statistics and the associated p-values for evaluating the same null hypotheses. The first set is shown to be asymptotically mutually independent, justifying the use of the Benjamini-Hochberg procedure. We further exploit the pairing structure through a two-step procedure aimed at improving power. Our theoretical results establish the key properties of the framework with respect to asymptotic FDR control and formally characterize the associated power gains of the two-step procedure. Importantly, our framework accommodates general dependence in the design matrix. Extensive simulations demonstrate that our methods outperform existing approaches – particularly those relying on empirical FDP estimates – in both power and FDR control accuracy, with notable gains in settings involving weaker signals, small sample sizes, or low target FDR levels.

Bias reduction in g-computation for covariate adjustment in randomized clinical trials

Xin Zhang

Pfizer Inc.

Abstract： G-computation is a powerful method for estimating unconditional treatment effects with covariate adjustment in randomized clinical trials. It typically relies on fitting canonical generalized linear models. However, this could be problematic when the sample size or event number is small relative to the number of covariates. Common issues include the underestimation of the variance and the potential nonexistence of maximum likelihood estimators. Bias reduction methods are commonly employed to address these issues, including Firth correction, which guarantees the existence of corresponding estimates. Yet, their application within g-computation remains underexplored. In this article, we analyze the asymptotic bias of g-computation estimators and propose a novel bias-reduction method that improves both estimation and inference. Our approach performs bias correction via generalized Oaxaca-Blinder estimators, and thus the resulting estimators are guaranteed to be bounded. The proposed debiased estimators use slightly modified versions of maximum likelihood or Firth correction estimators for nuisance parameters. Inspired by the proposed debiased estimators, we also introduce a simple small-sample bias adjustment for variance estimation, further improving finite-sample inference validity. Through extensive simulations, we demonstrate that our proposed method offers superior finite-sample performance, effectively addressing the bias-efficiency tradeoff. Finally, we illustrate its practical utility by reanalyzing a completed randomized clinical trial.

Higher-order debiased estimators of general treatment models

Zheng Zhang

Renmin University of China

Abstract： We have witnessed tremendous progress in developing the foundation for econometrics and causal inference in the past decades. The most popular paradigm in the current literature is the classical (first-order) semiparametric theory, in which a key building block is the (first-order) influence functions. However, it is now well known that estimators based on influence functions can be sub-optimal in terms of convergence rates in various settings. To address this issue, higher-order influence functions (HOIF) are developed, generalizing the classical semiparametric theory. However, most existing results in this regard focus on treatment effect parameters in explicit forms, such as average treatment effects (ATE). In applications, economists are often confronted with tasks of inferring more complex parameters, such as quantile treatment effects (QTE) or effects of complicated treatment regimes/policy. These more complex parameters can often only be implicitly defined as the solution to nonlinear estimating equations, which correspond to M/Z-estimation problems. Our current understanding of these problems is limited to the classical semiparametric theory. Given the foundational role of HOIF for estimating explicit parameters such as ATE, a modest step toward enriching the statistical foundation of econometrics and causal inference is to develop the corresponding higher-order estimators for those more complex parameters. To this end, we consider parameters of a class of non-separable structural models in the econometrics literature and develop a class of higher-order estimators for the target parameters. Statistical properties of these higher-order estimators are derived by leveraging recent advances in U-processes theory. Our proposed higher-order estimators relax complexity-reducing assumptions, quantified via Holder smoothness, on the nuisance parameters, compared to existing estimators for many important parameters in this class, including QTE and quantile dose-response functions, among others. Numerical experiments, including simulation studies and a real data analysis, are also conducted to corroborate our theoretical claims and illustrate how higher-order estimators can be used in practice.

Model-free Estimation of Latent Structure via Multiscale Nonparametric Maximum Likelihood

Ruiyi Yang

Shanghai Jiao Tong University

Abstract: Multivariate distributions often carry latent structures that are difficult to identify and estimate, and which better reflect the data generating mechanism than extrinsic structures exhibited simply by the raw data. In this talk, we propose a model-free approach for estimating such latent structures whenever they are present, without assuming they exist a priori. Given an arbitrary density, we construct a multiscale representation of the density and propose data-driven methods for selecting representative models that capture meaningful discrete structure. Our approach uses a nonparametric maximum likelihood estimator to estimate the latent structure at different scales and we further characterize their asymptotic limits. By carrying out such a multiscale analysis, we obtain coarseto-fine structures inherent in the original distribution, which are integrated via a model selection procedure to yield an interpretable discrete representation of it. As an application, we design a clustering algorithm based on the proposed procedure and demonstrate its effectiveness in capturing a wide range of latent structures.

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

Yujia Gu

Tsinghua University

Abstract: Randomized clinical trials are the gold standard for analyzing treatment effects. However, increasing costs and ethical concerns may limit trial recruitment, resulting in insufficient sample sizes and potentially invalid inference. Incorporating external trial data with similar characteristics (treatments, diseases, biomarkers, etc.) into the analysis appears promising for addressing these issues. Transfer learning, which in our context utilizes external trials as the source domain and current trials as the target domain, may offer a viable approach. 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. To cover broad types of randomized trials, we study this problem under stratified randomization, or more generally, covariate-adaptive randomization. For the algorithm, we adopt a parameter-based transfer learning approach to enhance the lasso-adjusted stratum-specific estimator developed for estimating the treatment effect. A key component in constructing the transfer learning estimator is deriving the regression coefficient estimation within each stratum, accounting for the bias between source and target data. To provide a theoretical foundation, we derive the convergence rate for the estimated regression coefficients and subsequently establish the asymptotic normality for 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 current trial data. Finally, we propose a consistent nonparametric variance estimator to facilitate inference that is robust to model misspecifications and applicable to various commonly used randomization procedures. Numerical studies demonstrate the effectiveness and robustness of our proposed estimator across various scenarios. Our study highlights the potential of transfer learning in analyzing randomized clinical trials.

Bias Correction for Semiparametric Regression Models

Stéphane Guerrier[1],Yanyuan Ma[2],Xuming He[3],Stéphane Guerrier[4]

1. Harvard University
2. Pennsylvania State University
3. Washington University in St. Louis
4. University of Geneva

Abstract： We consider a broad class of semiparametric regression models in which the conditional distribution of the response takes the form $f\{Y|\x\trans\bb+m(z),\phi\}$, which is known up to a parametric component $\bb$ of diverging dimension p, a smooth function $m(\cdot)$, and a dispersion parameter ϕ. Existing semiparametric literature on such models has primarily focused on semiparametric efficiency for $\bb$, typically treating ϕ and $m(\cdot)$ as nuisances and largely ignoring their finite-sample bias. However, the finite-sample bias of standard estimators can be substantial (especially when p is large relatively to n and/or dispersion is high) and can seriously undermine inference for $\bb$. Moreover, ϕ is often of direct scientific interest and requires accurate estimation. To address this gap, we propose SABRE, a simulation-based bias correction framework for this broad semiparametric model class. We establish asymptotic properties of SABRE for the subclass of generalized partially linear models, where bias reduction for $\bb$ and ϕ can be achieved without inflating variance, and we outline how the underlying principle may be adapted more generally. Comprehensive simulation studies and a real-data application on early-stage diabetes demonstrate the empirical effectiveness of SABRE in reducing bias and improving inference.

Studies in Label Shift

Yanyuan Ma

PSU

Abstract： In the context of discrete response label shift, we study the importance weights confidence set problem by a paradigm shift from traditional inversion-based inference to a direct matrix constraint framework. We use this framework to characterize a joint confidence region and extract marginal intervals via linear programming, deriving provably tighter bounds for importance weights while maintaining exact finite-sample validity. In the context of continuous response, we study the estimation and inference of a general target population characteristric bydeveloping doubly and singly robust estimators as well as the efficient estimator. Many ongoing and future developments will be discussed too. Applications will be discussed.

The implicit bootstrap: a percentile second-order correct interval estimation method

Mucyo Karemera[1],Samuel Orso[1],St éphane Guerrier[1],Maria-Pia Victoria-Feser[2],Min-ge Xie[3]

1. University of Geneva
2. University of Bologna
3. Rutgers University

Abstract： We explore the properties of an alternative to Efron's bootstrap confidence intervals in the form of a minimization problem. We show that, under general conditions, this method allows for the construction of asymptotically valid percentile confidence intervals of the parameter of interest. Furthermore, we show that these confidence intervals can also achieve second-order accuracy without the need of any additional correction terms or knowledge of the covariance matrix of an estimator. Thus, this approach combines the simplicity of Efron's percentile bootstrap with the accuracy of methods like the bootstrap-t or BCa. Simulation studies illustrate the comparative benefits of this method.

Partial optimality and applicability for multivariate equivalence testing

Luca Insolia[1],Yanyuan Ma[2],Younes Boulaguiem[1],St éphane Guerrier[1]

1. University of Geneva
2. The Pennsylvania State University

Abstract： Average equivalence testing aims at assessing whether an effect of interest, such as the difference in means between two experimental outcomes, lies within a predetermined region of practical equivalence. While univariate tests for normal means are widely used in modern applications, multivariate settings remain challenging, especially when experimental constraints limit the available sample sizes or introduce substantial noise. Established approaches, such as the Two One-Sided Tests (TOST), cannot adequately capture the complex multivariate nature of such data and lead to test procedures with limited power. To mitigate the conservativeness of TOST, we develop a corrected TOST framework by simultaneously adjusting its significance level and equivalence margins to ensure control of the test size and increase its power. We demonstrate that, in large samples, this approach leads to an optimal correction for the univariate TOST procedure, and we develop a further small-sample refinement to control its performance. In multivariate settings, where we show that an optimal correction does not exist, our proposal constrains marginal test sizes, leading to more balanced rejection regions, while maintaining overall size control and maximizing power in important cases. Through extensive simulation studies, we empirically demonstrate the superior performance of the proposed framework in scenarios relevant to real-world analyses. These scenarios include relatively small sample sizes, unknown and potentially heterogeneous variances, and various correlation structures. Finally, we illustrate the practical relevance of our approach with a case study related to the pharmaceutical sciences.

AI for dynamics network biology

Chunman Zuo

Sun Yat-sen University

Abstract： Understanding the dynamic principles of molecular networks, especially the critical regulation governing transitions of biological systems from stable to unstable states, is of fundamental importance for elucidating biological regulation. However, the averaging effect of global network modeling often masks weak and spatially heterogeneous critical signals, thereby limiting the discovery of regulatory mechanisms. To address this challenge, we developed a series of AI-based approaches from a new perspective that integrates microstructural domain identification with domain-specific network reconstruction. Specifically, we (i) proposed a complex network model that captures dynamic, directed, and heterogeneous intercellular interactions, avoiding over-smoothing of local structures and enabling accurate identification of microstructural domains; and (ii) established a path-independent heterogeneous network model that overcomes the bottleneck of adaptively decoupling cross-scale regulation, allowing prediction of key regulatory axes that drive critical-state transitions. Using this framework, we further identified and experimentally validated key targets involved in gastric cancer initiation. These results provide a new strategy for dissecting critical regulatory mechanisms in complex biological systems.

Local-and-Global Information-Preserving Statistical Manifold Learning for Single-Cell Transcriptomics

Hongyi Xin
Shanghai Jiao Tong University

Abstract： Geometry-preserving dimension reduction is critical for single-cell transcriptomics, where low-dimensional distances should reflect biological divergence between cell types along the transcriptomic manifold. Due to inadequate metrics, the global structure is not sufficiently preserved in the low-dimensional manifold in standard dimension reduction regimes. We model RNA counts as Multinomial samples, leveraging their hierarchical closure property: gene-level counts refine functional gene-group counts via nested Multinomial distributions. Extending Chentsov's Theorem, we show that the Fisher-Rao metric on coarse (gene-group) and fine (gene) statistical manifolds is isometric. Following this isometry property, we propose InfoGlobe, an information-preserving statistical manifold learning framework that projects cells from high-dimensional hyperspheres (full transcriptome) to low-dimensional hyperspheres (functional groups) while preserving information geometry. Embeddings on the low-dimensional sphere explicitly represent Multinomial distributions by functional gene groups. Benchmarks demonstrate superior preservation of local-and-global cell-type geodesic distances, automatic and robust gene-group discovery, nuanced cell subtype resolution without manual feature engineering and natural batch effect mitigation without explicit alignments.

Deciphering Tissue Spatial Heterogeneity: Methods for the Identification of Spatially Variable Genes and the Characterization of Spatial Structures

Xin Yuan
Shanghai Jiao Tong University

Abstract： The identification of spatially variable genes (SVGs) and the characterization of spatial structures are two central objectives in spatial transcriptomics (ST) analysis. However, the high dimensionality, spatial dependency, and rapidly increasing scale and resolution of ST data pose significant statistical and computational challenges to achieving these goals.

This study introduces two statistical methodologies to address these problems. HEARTSVG is a nonparametric, test-based framework for SVG detection that circumvents distributional assumptions and obviates the need for predefined spatial patterns. Simulation studies and empirical evaluations across twelve diverse ST datasets demonstrate that HEARTSVG achieves superior discriminative accuracy (average Score=0.948), improved computational efficiency, and effective control of false discovery, outperforming existing state-of-the-art approaches.

To facilitate spatial domain inference in ultra-large ST datasets, HERGAST is proposed—an attention-based graph learning framework embedded within a novel Divide-Iterate-Conquer (DIC) strategy. By constructing a heterogeneous graph that simultaneously models local spatial adjacency and global transcriptomic similarity, HERGAST enables scalable inference while mitigating oversmoothing artifacts commonly introduced by data partitioning. Extensive simulations reveal a consistent >10% improvement in adjusted Rand index (ARI) over competing methods. Applications to human colorectal and breast cancer datasets highlight HERGAST's ability to uncover biologically meaningful spatial structures, including tumor-enriched $SPP1^{+}$ macrophage clusters and spatially distinct oncogene expression domains.

Collectively, these methods provide statistically principled and computationally scalable solutions for spatial feature selection and domain reconstruction in large-scale ST data, facilitating interpretable and robust biological inference.

Meta-Encoder: Integrating Multiple Pathological Foundation Models to Enhance Predictive Accuracy for Key Cancer Biomarkers and Spatial Omics

Ruitian Gao

Shanghai Jiao Tong University

Abstract： The emergence of diverse pathological foundation models has empowered computational pathology tasks within the field of oncology, including tumor subtyping, cancer prognosis, biomarker prediction, gene expression prediction, and so on. However, variations in model architecture and data sources hinder consistent downstream performance and complicate centralized training. Specifically, the lack of data sharing makes retraining foundation models with pooled data infeasible. Here, we propose the Meta-Encoder, a unified framework that integrates features from multiple pathological foundation models to generate a comprehensive representation, achieving superior performance in downstream cancer detection tasks compared to single foundation models. While single models suffice for low-complexity univariate tasks such as cancer diagnosis and prognosis, the Meta-Encoder consistently rivals the best-performing single model, alleviating concerns over model selection. For high-dimensional tasks such as multiplex protein and gene expression prediction within tumor tissues, the attention-based strategies of our Meta-Encoder framework demonstrate substantial advantages over single models, offering an optimal balance of performance and efficiency. By harnessing the complementary strengths of multiple foundation models, the Meta-Encoder enhance the molecular characterization of pathology images, thereby advancing precision oncology.

Note