INTERNATIONAL CHINESE STATISTICAL ASSOCIATION
泛華統計協會

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

2026 泛华统计协会中国会议
ICSA China Conference

Shenzhen, Guangdong, China
June 27- June 29, 2026



International Chinese Statistical Association

China Conference 2026

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 27 - June 29, 2026
Southern University of Science and Technology
Shenzhen, Guangdong, China

Organized by
International Chinese Statistical Association

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

©2026
International Chinese Statistical Association

Content

Welcome & Conference Essentials 1

2026 ICSA China Conference 1

ICSA Core Members 2

Conference Committees 3

Conference Guide 7

Check-in, Registration 7

Reception 8

Conference Venue 9

Shuttle Bus 15

Meals 19

Scientific Program 20

Program Overview 20

Keynote Lectures 21

Junior Researcher Award Winners 25

Poster Session 26

Short Course 28

Parallel Invited Sessions Schedule 29

a.June 27th PM (14:00-15:40) 29

b.June 27th PM (16:00-17:40) 30

c. June 28th PM (14:00-15:40) 31

d. June 28th PM (16:00-17:40) 32

e. June 29th AM (08:30-10:10) 34

f. June 29th AM (10:30-12:10) 35

Detailed Invited Sessions Schedule 37

a.June 27th PM (14:00-15:40) 37

b. June 27th PM (16:00-17:40) 43

c. June 28th PM (14:00-15:40) 50

d. June 28th PM (16:00-17:40) 56

e. June 29th AM (08:30-10:10) 63

f. June 29th AM (10:30-12:10) 69

Abstracts of Invited Sessions 77

S007: Advanced Statistical Modeling and Integration of Complex Data 77

S010: Advances in High-Dimensional and Nonparametric Statistics 78

S013: Advances in Modeling Frameworks for Structured and Complex Data 79

S023: Advances in Time Series Modeling and Functional Data Analysis 80

S028: Bayesian Regression and Variable Selection for Structured and High-Dimensional Data 81

S031: Bridging Statistics and Modern AI: Foundations for Deep Learning and Generative Models 82

S033: Computational and Statistical Advances in Association Modeling 83
S036: Data Integration Under Different Data Types: New Methodology Developments 84
S038: Emerging Statistical Frameworks for High-Dimensional and Biased Data 85
S039: Estimation and Testing in Sparse, Quantile, and Survival Models 86
S042: From Cells to Tissues: Statistical Frameworks for Spatial Omics 86
S055: Innovations in Regression and Change Detection for Complex Data 88
S073: Modern Statistical Methods for Complex Data: Hypothesis Testing, Estimation and Application 89
S088: Precision Medicine and High-Dimensional Data Inference 90
S097: Recent Advances in Data-Driven Decision Making 91
S104: Recent Advances in Omics Data Analysis Techniques 92
S109: Recent Advances in Statistical Designs for Early-Phase Oncology Clinical Trials 93
S114: Recent Advances on the Analysis of Complex Data 94
S126: Statistical and Computational Methods for Omics and Biological Discovery 95
S128: Statistical Foundations for Machine Learning and AI 96
S050: Frontiers of Transfer Learning, Data Fusion, and Adaptive Evidence Generation: Theory, Methods, and Practice 97
S136: Statistical Learning and Modeling for Complex Data 98
S147: Study Design and Joint Modeling with Longitudinal and Survival Data 99
S154: When Large Language Model Meets Statistics 100
S005: Advanced Statistical Methods for Complex Survival Data 101
S021: Advances in Statistical Modeling for Causal Inference and High-dimensional Complex Data 102
S025: AI and Big Data Analysis 103
S030: Bootstrap and Model Averaging Methods for Robust Statistical Learning 105
S041: Advances in Statistical Inference for Dependent and High-Dimensional Data 106
S045: Frontiers in Biostatistics: Robust Inference and Machine Learning 107
S129: Statistical Foundations for Preference Learning, Representation, and Heterogeneity in the Era of AI 108
S054: Inference and Learning Under Heterogeneous Data 109
S056: Innovative Approaches in Machine Learning and Health Data Analytic 110
S058: Innovative Statistical Design Methods for Modern Clinical Trials 111
S059: Integrative Statistical Methods for Genetic Association and Multi-Omics Studies 112
S075: Modern Statistical Methods for High-Dimensional and Structured Data 113
S076: Modern Statistical Methods for High-Dimensional Data and Machine Learning Models 114
S079: New Advances in Infectious Disease Modeling 115
S107: Recent Advances in Spatial Omics and Single-cell Data Analysis 116
S112: Recent Advances of Time Series Analysis and Change-Point Detection 117
S117: Recent Developments on Applied Statistics 118
S122: Some New Methods of Robust Statistical Learning for Complex Data 119
S124: Statistical Analysis of Complex Data 120
S127: Statistical and Computational Methods for Single-Cell, Spatial, and Multi-Omics Integration 121
S130: Statistical Foundations of LLM Behavior, Alignment, and Automated Reasoning 122

S131: Statistical Frontiers in Spatial Omics 123
S140: Statistical Methods and Applications 124
S148: Survival Analysis and Related Statistical Methods 126
S106: Recent Advances in Sequential Methodologies 127
S012: Advances in Machine Learning Methods for Survival and Event-Time Data 128
S002: Advanced Methods for Complex Data Analysis 129
S092: Recent Advances for the Analysis of Complex Data 130
S014: Advances in Modern High-dimensional Statistics 130
S022: Advances in Survival Analysis and Causal Inference 131
S019: Advances in Statistical Methods for Clinical Trials: Longitudinal Data and Complex Data Mechanisms 132
S029: Big Data and AI 134
S040: Experimental Design and Optimization 134
S052: High-Dimensional Inference and P-Value Combination 136
S062: Machine Learning and Statistical Modeling Applications 136
S063: Microbiome Intelligence: Large Models, Big Data, and Statistical Advances 137
S069: Modern Statistical Inference and Modeling for Complex and Structured Data 138
S070: Modern Statistical Learning and High Dimensional Modeling 139
S078: New Advances in Differential Privacy 140
S082: New Developments of Statistical Methods in Bio-Medical Studies 141
S086: Novel Statistical and Machine Learning Methods for Analyzing Omics Data 142
S099: Recent Advances in High-Dimensional and Functional Hypothesis Testing 143
S100: Recent Advances in High-Dimensional Data Analysis 144
S115: Recent Advances on the Analysis of Failure Time Data 145
S121: Scalable and Distributed Statistical Learning for Complex Data Structures 146
S135: Statistical Innovations for Spatial Transcriptomics and 3D Genomics 147
S139: Statistical Learning with AI 148
S146: Structured Data Analysis and Dependence Modeling 149
S150: The Frontier Research of Statistical Process Quality Control under AI era 150
S144: Statistics and Machine Learning at the Frontiers of Genomics and Genetics 151
S001: Advances in Statistical Methods for Genetic and Genomic Studies 152
S003: Advanced Methods for Dependence, Signals and Spatio-Temporal Detection 154
S015: Advances in Predictive and Integrative Modeling for Complex Health Data 155
S016: Advances in Statistical Inference for Complex and High-Dimensional Models 156
S026: AI-Driven Genomics and Precision Medicine: From Prediction to Personalization 157
S024: Advancing Complex Dependent Data Analysis 158
S034: Conditional Modeling and Causality: Generative Models for Learning and Integration 159
S047: Frontiers in Discrete-Valued Time Series Modeling 160
S048: Frontiers in Learning and Optimization: From GenAI to Dynamic Operations 161
S049: Frontiers in Methods for Complex-structured Survival Data 162
S060: Interval-Valued Data Modelling and Forecasting 163

S066: Modern Methods for Omics Inference and Clinical Translation 164
S068: Modern Perspectives on High-Dimensional Statistical Theory and Methodology 165
S071: Modern Statistical Learning for Complex Data Analysis 166
S084: New Methods for the Analysis of Complex Data 167
S085: New Statistical and Coputational Methods for Large Data Analysis 168
S096: Recent Advances in Computational Methods for Single-Cell and Spatial Omics..... 169
S103: Recent Advances in Network Analysis 170
S118: Recent Progress in Dependent Data Analysis 171
S133: Statistical Inference for Complex Data 172
S141: Statistical Methods for Genetics and Genomics 173
S142: Statistical Methods for Reliable Machine Learning..... 174
S143: Statistical Modeling and Inference for Large Language Models..... 176
S149: The Broad Roles of Design in Modern Statistical Analysis..... 177
S153: Uncertainty Quantification in Modern Statistics and Machine Learning 178
S155: Matrix Data..... 179
S032: Causal Learning in Modern Health and Survey Data 180
S011: Advances in High-Dimensional Dependence Modeling..... 181
S018: Advances in Statistical Machine Learning Methods for Complex Data..... 182
S037: Decoding Complexity: Genomic Diversity, Disease Mechanisms, and Evolutionary Insights 183
S043: From Data to Biomarkers: Statistical Innovation for Complex Challenges in Modern Genetics, Genomics and Proteomics 184
S044: From Probabilistic Theory to Causal Inference in Complex Systems..... 186
S051: High Dimensional Data Analysis and Change Point Inference 186
S064: Model Averaging and Related Topics..... 187
S065: Model Averaging for Time Series Data..... 188
S072: Modern Statistical Methods and Artificial Intelligence: From Theory to Biomedical Applications. 189
S077: Mordern Distributional Methods for Learning, Inference, and Robustness..... 190
S081: New Developments in Survey Sampling 192
S083: New Frontiers in Spatial Modeling: Dissimilarity, Extremes, and High-Dimensional Clustering ... 193
S087: Post and Semi-Supervised Inference..... 194
S090: Recent Advance in Statistical Learning 195
S091: Recent Advances for Statistical Analysis of Censored Data..... 196
S093: Recent Advances in Complex Data Analysis and New Applications 197
S098: Recent Advances in Designs and Statistical Methods for Clinical Trials Research 198
S105: Recent Advances in Selective Inference, Data Assimilation, and Functional Data..... 198
S111: Recent Advances in Statistical Methods for Genetic Data..... 199
S119: Regression Modeling and Efficient Estimation for Complex Data..... 200
S132: Statistical Inference and Subgroup Analysis for Specialized Data Types 201
S137: Statistical Learning for Complex Data..... 203
S145: Statistics, Machine Learning, and AI 204
S057: Innovative Machine Learning and Statistical Methods for Complex Biomedical Data..... 205

S151: Time Series Econometrics and Applications 206
S004: Advanced Statistical and Machine Learning Methods in Genetics and Genomics 206
S008: Advances in Bayesian Modeling for Big and Complex Data 207
S009: Advances in High-Dimensional and Complex Data Analysis 208
S017: Advances in Statistical Learning and AI for Biological and Health Sciences 209
S020: Advances in Statistical Methods for Omics and Health Data Analysis 210
S027: Applications in Biomedical Sciences 212
S046: Frontiers in Causal Inference: Addressing Time-Dependence, Fairness, and Long-Term Effects in Complex Data 213
S053: High-Dimensional Statistics 214
S067: Modern Mixture Model Methodology: Advances in Modeling, Inference, and Applications 215
S080: New Advances in Microbiome Data Analysis and Related Fields 216
S089: Random Matrix Theory and Applications 217
S006: Advanced Statistical Methods for Network Data and Latent Variable Models 218
S094: Recent Advances in AI for Biological Sciences 219
S095: Recent Advances in Complex Data Analysis 219
S101: Recent Advances in Longitudinal Data Analysis 220
S102: Recent Advances in Modeling Survival Data 221
S108: Recent Advances in Spatio-Temporal Data Analysis 222
S110: Recent Advances in Statistical Learning and Big Data 223
S113: Recent Advances on Statistical Inference and Econometrics 224
S116: Recent Advances in Complex Time Series and Dynamic Network Analysis 225
S120: Robust and Flexible Statistical Methods for Complex and High-Dimensional Data 226
S125: Statistical and Computational Advances for Single-Cell Genomics 227
S134: Statistical Inference for Complex Data Structures 228
S152: Toward Efficient Clinical Trials: Statistical Methods That Leverage Baseline Covariates 229
S074: Modern Statistical Methods for Complex Longitudinal Data and Causal Inference 230

Southern University of Science and Technology 232
School of Science, SUSTech 235
Department of Statistics and Data Science, SUSTech 237
Faculty Position Announcement 239

Logistics & Practical Information 241
Accommodation 241
Transportation Guide 243
Medical Information 255
Weather Notice 259
Local Attractions 261
24-Hour Food & Beverage Establishments 268

Welcome & Conference Essentials

2026 ICSA China Conference

June 27- June 29, Shenzhen , Guangdong , China

Welcome to the International Chinese Statistical Association (ICSA) 2026 China Conference!

The 2026 ICSA China Conference will be held in Shenzhen, Guangdong, China, from June 26 to June 29, 2026. It is co-organized by ICSA and Southern University of Science and Technology (SUSTech). The conference venue is located at the Convention Center and Chi Wah Lecture Hall (the Third Lecture Hall), School of Business, Southern University of Science and Technology (SUSTech). The theme of this conference is "**From Data to Decisions: Statistical Innovation for Complex Challenges**", aiming to explore cutting-edge developments, interdisciplinary collaboration, and innovative applications of statistics and data science in the era of AI and big data.

The executive and organization committees have been working diligently to put together a robust and comprehensive program, including four keynote lectures, 151 invited sessions, 21 poster presentations. The four excellent keynote lectures will be given by ***Dr. Jeremy Taylor*** (University of Michigan, USA), ***Dr. Annie Qu*** (University of California, Santa Barbara, USA), ***Dr. Niansheng Tang*** (Yunnan University, China), and ***Dr. Weijie Su***(University of Pennsylvania, USA), who are the top experts in their fields.

With your full support, this conference will attract more than 600 statisticians and data scientists working in academia, government, and industry from all over the world. We hope that the conference provides great opportunities for learning, networking, collaborations and recruiting. You will share thoughts and ideas with conference guests, receive inspiration from existing research and develop new directions. The local organizing committee and more than 80 student volunteers from SUSTech have made great efforts to arrange the meeting logistics and social events, including the evening banquet on Sunday, June 28. We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

As a major innovative metropolis in southern China, Shenzhen enjoys a subtropical climate with convenient transportation, well-connected by high-speed rail, ports, and international airports. It features a vibrant modern atmosphere, high-quality business and conference facilities, diverse cuisines including authentic Cantonese food and fresh seafood, and convenient urban services. We welcome you to enjoy academic exchanges and experience the unique charm of this dynamic coastal city.

Thanks for coming to the 2026 ICSA China Conference in Shenzhen!

Jianguo (Tony) Sun and George C. Tseng, on behalf of

2026 ICSA China Conference Executive and Program Committees

ICSA Core Members

EXECUTIVES

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

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

President-elect: Sammi Tang (Sammi.tang@astellas.com)

Executive Director: Jun Zhao (executive.director@icsa.org)

ICSA Treasurer: Xin He (treasurer@icsa.org)

The ICSA Office Manager: Grace Ying Li (office@icsa.org)

BOARD of DIRECTORS

Xiaotian Chen (xiaotian.chen@abbvie.com)

Yichuan Zhao (yichuan@gsu.edu)

Dehan Kong (dehan.kong@utoronto.ca)

Grace Shieh (gshieh2@stat.sinica.edu.tw)

Jingjing Ye (Jingjing.ye@beonemeds.com)

Yuguo Chen (yuguo@illinois.edu)

Haitao Chu (hchu.xy@gmail.com)

Yuanjia Wang (yw2016@cumc.columbia.edu)

Xinping Cui (xpcui@ucr.edu)

Yushi Liu (liu_yushi@lilly.com)

Jialiang Li (jialiang@nus.edu.sg)

George C. Tseng (ctseng@pitt.edu)

Kun Chen (kun.chen@uconn.edu)

Song Yang (yangso@nhlbi.nih.gov)

Jianchang Lin (Jianchang.Lin@takeda.com)

Rong Chen (rongchen@stat.rutgers.edu)

Hongyu Zhao (hongyu.zhao@yale.edu)

Sammi Tang (Sammi.tang@astellas.com)

Conference Committees

Executive Committee

Rong Chen (Co-Chair)	Rutgers, The State University of New Jersey
Qiman Shao (Co-Chair)	Southern University of Science and Technology
Jianguo Sun	Southern University of Science and Technology
George C. Tseng	University of Pittsburgh
Jun Zhao	AbbVie

Scientific Program Committee

Jianguo Sun (Co-Chair)	Southern University of Science and Technology
George C. Tseng (Co-Chair)	University of Pittsburgh
Zhanrui Cai	The University of Hong Kong
Jade (Huanyu) Chen	Fortvita Biologics
Kun Chen	University of Connecticut
Xuerong Chen	Southwestern University of Finance and Economics
Yong Chen	University of Pennsylvania
Mingyue Du	Jilin University
Chiungyu Huang	University of California, San Francisco
Jian Huang	The Hong Kong Polytechnic University
Tao Hu	Capital Normal University
Ching-Kang Ing	Tsinghua University, Taiwan
Hongkai Ji	Johns Hopkins University
Jiancheng Jiang	Great Bay University
Xuejun Jiang	Southern University of Science and Technology
Chen-hung Kao	Academia Sinica, Taiwan
Meiling Lee	University of Maryland
Changcheng Li	Dalian University of Technology
Gang Li	University of California, Los Angeles
Huiqiong Li	Yunnan University
Jessica Li	Fred Hutchinson Cancer Center
Jialiang Li	National University of Singapore

Lexin Li	University of California, Berkeley
Huazhen Lin	Southwestern University of Finance and Economics
Shili Lin	The Ohio State University
Tsung-I Lin	Chung Hsing University
Qizhai Li	Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Quefeng Li	University of North Carolina at Chapel Hill
Shuwei Li	Guangzhou University
Ching-Ti Liu	Boston University
Jingyuan Liu	Xiamen University
Yanyan Liu	Wuhan University
Zhigang Li	University of Florida
Tianzhou Ma	University of Maryland
Li-Xuan Qin	Memorial Sloan Kettering Cancer Center
Xinyuan Song	The Chinese University of Hong Kong
Fasheng Sun	Northeast Normal University
Liuquan Sun	Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Yuying Sun	Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Weijie Su	University of Pennsylvania
Lu Tang	University of Pittsburgh
Guoliang Tian	Southern University of Science and Technology
Lu Tian	Stanford University
Chunjie Wang	Changchun University of Technology
Dehui Wang	Liaoning University
Lu Wang	University of Michigan
Ming Wang	Case Western Reserve University
Peijie Wang	Jilin University
Tianying Wang	Colorado State University
Xueqin Wang	University of Science and Technology of China
Yuanjia Wang	Columbia University
Zhaojun Wang	Nankai University
Yingying Wei	The Chinese University of Hong Kong
Zheyang Wu	Worcester Polytechnic Institute
Fang Yao	Peking University

Fengting Yi	Yunnan University
Tianwei Yu	The Chinese University of Hong Kong, Shenzhen
Anru Zhang	Duke University
Tingting Zhang	University of Pittsburgh
Xinyu Zhang	Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Hui Zhao	Zhongnan University of Economics and Law
Shishun Zhao	Jilin University
Xingqiu Zhao	The Hong Kong Polytechnic University
Yichuan Zhao	Georgia State University
Shurong Zheng	Northeast Normal University
Shouhao Zhou	The Pennsylvania State University
Xiang Zhou	Yale University
Yong Zhou	East China Normal University
Fukang Zhu	Jilin University
Zhongyi Zhu	Fudan University
Changliang Zou	Nankai University
Guohua Zou	Capital Normal University

Junior Researcher Award Committee

Bin Zhang (Chair)	Cincinnati Children's Hospital Medical Center
Xuejun Jiang	Southern University of Science and Technology
Dayu Sun	Indiana University

IT and Website Committee

Chengsheng Jiang (Co-Chair)	ICSA
Anqi Song (Co-Chair)	Southern University of Science and Technology
Shengjie Zhu	Southern University of Science and Technology

Local Organization Committee

Qiman Shao (Co-Chair)	Southern University of Science and Technology
Xuejun Jiang (Co-Chair)	Southern University of Science and Technology
GuanHua Chen	Southern University of Science and Technology
Xin Chen	Southern University of Science and Technology

Yanqing Hu	Southern University of Science and Technology
Xiyun Jiao	Southern University of Science and Technology
Fang Kong	Southern University of Science and Technology
Ting Li	Southern University of Science and Technology
Zeng Li	Southern University of Science and Technology
Yifang Ma	Southern University of Science and Technology
Jian Qing Shi	Southern University of Science and Technology
Jianguo Sun	Southern University of Science and Technology
Yuxin Tao	Southern University of Science and Technology
Guoliang Tian	Southern University of Science and Technology
Chao Wang	Southern University of Science and Technology
Hongxin Wei	Southern University of Science and Technology
Haolei Weng	Southern University of Science and Technology
Lili Yang	Southern University of Science and Technology
Linyi Yang	Southern University of Science and Technology
Peng Yang	Southern University of Science and Technology
Haoran Zhang	Southern University of Science and Technology
Zhuosong Zhang	Southern University of Science and Technology

Poster Session Committee

Yuxin Tao (Chair)	Southern University of Science and Technology
Catherine Liu	The Hong Kong Polytechnic University
Haolei Weng	Southern University of Science and Technology
Jun Yu	Beijing Institute of Technology
Xiang Zhan	Southeast University
Anru Zhang	Duke University
Rongmao Zhang	Zhejiang Gongshang University

Short Course Committee

Zhixiang Lin (Chair)	The Chinese University of Hong Kong
Ben Dai	The Chinese University of Hong Kong
Cong Li	Takeda pharmaceuticals, Boston, USA
Fangda Song	Chinese University of Hong Kong, Shenzhen

Check-in, Registration

1.Conference Venue: Southern University of Science and Technology (SUSTech)

2.Check-in Location and Time:

(1) June 26 (14:00 - 22:00): Dacheng Auditorium (Lecture Hall 114), Business School Building, SUSTech

(2) June 27 (08:00 - 12:00): 1st Floor, Conference Center, SUSTech

(13:30 - 17:00): 1st Floor, Chi Wah Lecture Hall (Third Teaching Building)

(3) June 28 (09:00 - 12:00): 1st Floor, Conference Center, SUSTech

(13:30 - 17:00): 1st Floor, Chi Wah Lecture Hall (Third Teaching Building)

(4) June 29 (08:00 - 12:00): 1st Floor, Chi Wah Lecture Hall (Third Teaching Building)

3.Notes:

(1) On-site registration is conducted in separate zones. All guests are kindly reminded to confirm their payment status beforehand: fully paid guests may proceed to the corresponding zones sorted by the initial letter of their surnames for check-in and registration; guests with outstanding payments shall first complete payment at the on-site payment area before registration. Please follow the guidance of on-site staff and complete all formalities in an orderly manner.

(2) Complimentary shuttle buses between the hotels and SUSTech will be available during the registration period on June 26. For detailed schedules, please refer to the conference transportation arrangements.

会议签到、注册

1. **会议地点：**南方科技大学

2. **会议签到、注册时间及地点：**

（1）6 月 26 日（14:00–22:00）：南方科技大学商学院大楼 114 大成礼堂；

（2）6 月 27 日（08:00–12:00）：南科大会议中心 1 楼；

（13:30–17:00）：智华教学楼（第三教学楼）1 楼；

（3）6 月 28 日（09:00–12:00）：南科大会议中心 1 楼；

（13:30–17:00）：智华教学楼（第三教学楼）1 楼；

（4）6 月 29 日（08:00–12:00）：智华教学楼（第三教学楼）1 楼。

备注：

1. 现场注册设置分区办理，请各位嘉宾先行确认自身缴费状态：已完成缴费按姓氏首字母前往对应区域办理签到注册；未缴费嘉宾前往现场缴费区办理缴费注册，全体人员请按照现场工作人员指引有序办理。

2.6 月 26 日注册时段内设有酒店至南科大往返专属接驳班车，具体班次详见会议交通安排。

邀请函
INVITATION

Reception and Promotional Introduction Session of the Department of Statistics and Data Science

Time June 26, 20:00–22:00

Venue Dacheng Auditorium (Lecture Hall 114), Business School Building, SUSTech

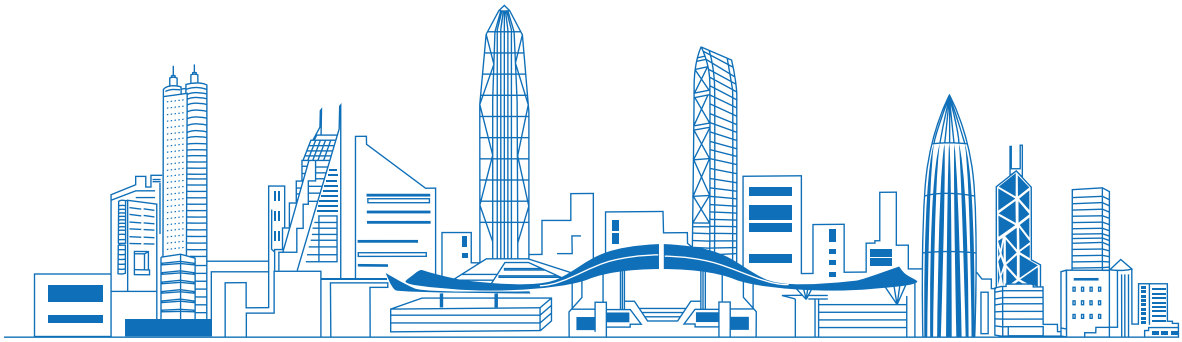
(Same location as check-in and registration)

统计与数据科学系宣传接待会

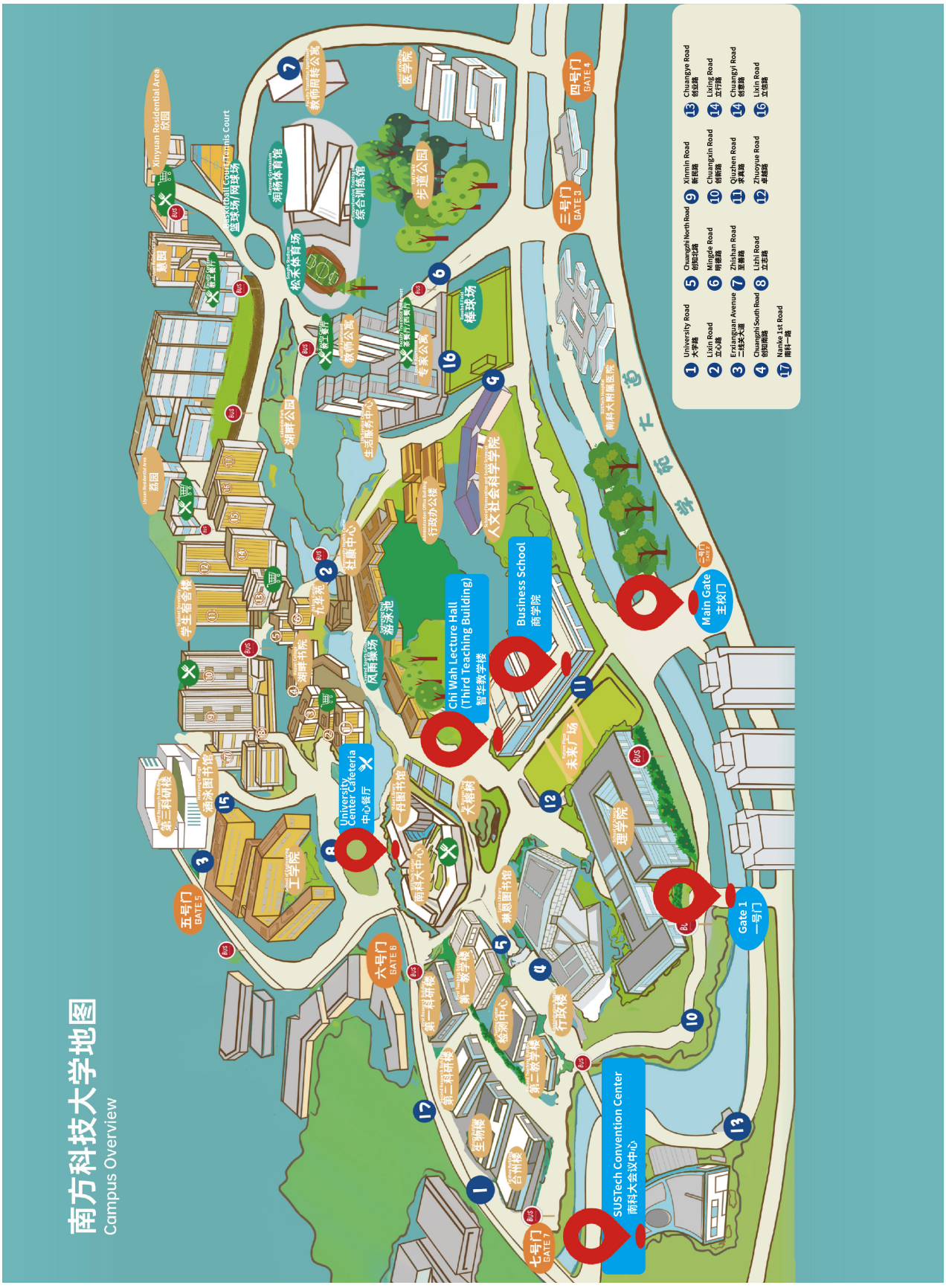
时间 6月26日20:00-22:00

地点 南方科技大学商学院大成礼堂114报告厅

(和注册同样的地方)



Conference Venue







June 26: Check-in & Registration; Reception; Poster Evaluation Dacheng Auditorium (Lecture Hall 114), Business School Building, SUSTech

6 月 26 日签到、注册；统计与数据科学系宣传接待会；Poster 评选：南科大商学院大成礼堂 114 报告厅

1F, Business School Building

June 27 AM and June 28 AM: Keynote lectures

Concert Hall, 1/F of Convention Center, SUSTech and Round Table Meeting Room, 4/F of Convention Center, SUSTech respectively

6 月 27 日上午、6 月 28 日上午大会报告
会议中心 1 楼音乐厅、4 楼圆形会议室







1F, Convention Center, SUStech



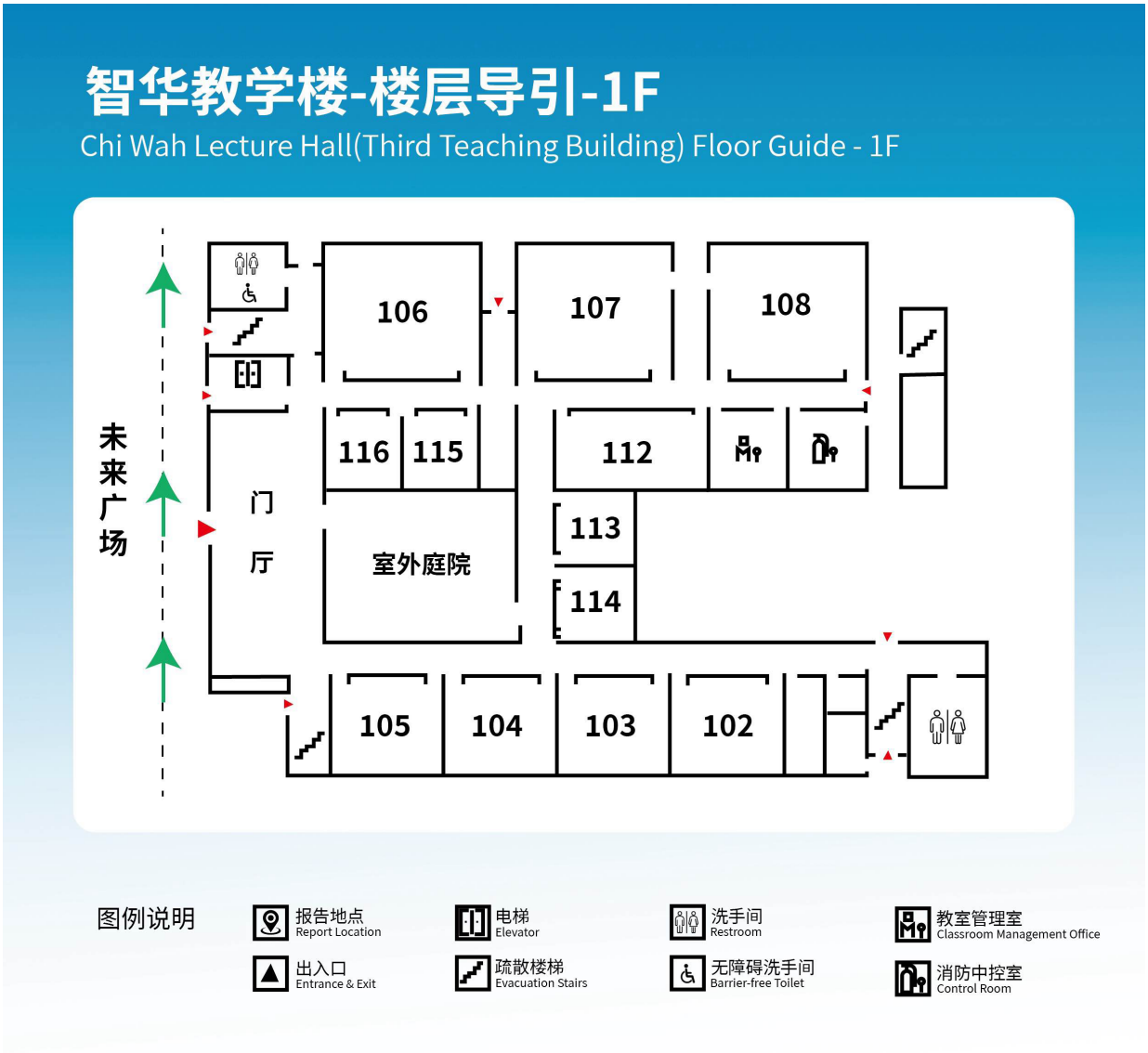
4F, Convention Center， SUSTech

June 27 PM, June 28 PM, June 29 AM: Parallel invited sessions

Chi Wah Lecture Hall (Third Teaching Building)

6 月 27 日下午、6 月 28 日下午、6 月 29 日上午分会场报告

智华教学楼（第三教学楼）



1F, Chi Wah Lecture Hall(Third Teaching Building)

June 27 PM, June 28 PM, June 29 AM: Parallel invited sessions

Chi Wah Lecture Hall (Third Teaching Building)

6 月 27 日下午、6 月 28 日下午、6 月 29 日上午分会场报告

智华教学楼（第三教学楼）



2F, Chi Wah Lecture Hall(Third Teaching Building)

June 27 PM, June 28 PM, June 29 AM: Parallel invited sessions

Chi Wah Lecture Hall (Third Teaching Building)

6 月 27 日下午、6 月 28 日下午、6 月 29 日上午分会场报告

智华教学楼（第三教学楼）



3F, Chi Wah Lecture Hall(Third Teaching Building)

Shuttle Bus

Shuttle Bus Schedule

Complimentary conference shuttle buses will be arranged between designated hotels and SUSTech venues during the conference period. Detailed timetable is listed below:

Date	Departure Time	Route
June 26	14:00–22:30	Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch) → Business School Building, SUSTech **Departures from the Hotels:** 14:00, 15:00, 16:00, 17:00, 18:00, 19:00, 20:00, 21:00 Business School Building, SUSTech →Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch) **Departures from SUSTech:** 14:30, 15:30, 16:30, 17:30, 18:30, 19:30, 20:30, 21:30, 22:30
	14:00–22:30	Sheraton Shenzhen Nanshan→ Business School Building, SUSTech **Departures from the Hotels:** 14:00, 15:00, 16:00, 17:00, 18:00, 19:00, 20:00, 21:00 Business School Building, SUSTech→Sheraton Shenzhen Nanshan **Departures from SUSTech:** 14:30, 15:30, 16:30, 17:30, 18:30, 19:30, 20:30, 21:30, 22:30
	14:00–22:30	Vienna Good Sleep International Hotel (Tanglang Metro Station) →SUSTech Business School Building **Departures from the Hotels:** 14:00, 15:00, 16:00, 17:00, 18:00, 19:00, 20:00, 21:00 Business School Building, SUSTech→Vienna Good Sleep International Hotel (Tanglang Metro Station) **Departures from SUSTech:** 14:30, 15:30, 16:30, 17:30, 18:30, 19:30, 20:30, 21:30, 22:30
June 27	Departs at 08:10	Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch) → SUSTech Convention Center
	Departs at 08:10	Sheraton Shenzhen Nanshan → Convention Center, SUSTech
	Departs at 08:10	Vienna Good Sleep International Hotel (Tanglang Metro Station) → Convention Center, SUSTech
	Departs at 18:00	Business School Building, SUSTech→ Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch)
	Departs at 18:00	Business School Building, SUSTech→ Sheraton Shenzhen Nanshan
	Departs at 18:00	Business School Building, SUSTech→ Vienna Good Sleep International Hotel (Tanglang Metro Station)

接驳班车

June 28	Departs at 09:00	Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch) → Convention Center, SUSTech
	Departs at 09:00	Sheraton Shenzhen Nanshan →Convention Center, SUSTech
	Departs at 09:00	Vienna Good Sleep International Hotel (Tanglang Metro Station) → Convention Center, SUSTech
	Departs at 18:00	Business School Building, SUSTech→ Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch)
	Departs at 18:00	Business School Building, SUSTech→ Sheraton Shenzhen Nanshan
	Departs at 18:00	Business School Building, SUSTech→ Vienna Good Sleep International Hotel (Tanglang Metro Station)
	18:00 (Banquet outbound)	Business School Building, SUSTech → Sheraton Shenzhen Nanshan
	21:00 (Banquet return trip)	Sheraton Shenzhen Nanshan → respective hotels
June 29	Departs at 07:50	Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch) → Convention Center, SUSTech
	Departs at 08:00	Sheraton Shenzhen Nanshan → Convention Center, SUSTech
	Departs at 08:00	Vienna Good Sleep International Hotel (Tanglang Metro Station) → Convention Center, SUSTech
	Departs at 12:20	Business School Building, SUSTech→ Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch)
	Departs at 12:20	Business School Building, SUSTech→ Sheraton Shenzhen Nanshan
	Departs at 12:20	Business School Building, SUSTech → Vienna Good Sleep International Hotel (Tanglang Metro Station)

接驳班车时刻表

会议期间组委会免费提供接驳班车，往返合作酒店与南科大各会场，具体发车时刻表如下：

乘车时间		乘车地点
6 月 26 日	14:00-22:30	璟峯酒店→南科大商学院 发车时间点：14:00,15:00,16:00,17:00,18:00,19:00,20:00,21:00 南科大商学院→璟峯酒店 发车时间点：14:30,15:30,16:30,17:30,18:30,19:30,20:30,21:30，22:30
	14:00-22:30	博林天瑞喜来登酒店→南科大商学院 发车时间点：14:00,15:00,16:00,17:00,18:00,19:00,20:00,21:00 南科大商学院→博林天瑞喜来登酒店 发车时间点：14:30,15:30,16:30,17:30,18:30,19:30,20:30,21:30,22:30
	14:00-22:30	维也纳好眠国际酒店→南科大商学院 发车时间点：14:00,15:00,16:00,17:00,18:00,19:00,20:00,21:00 南科大商学院→维也纳好眠国际酒店 发车时间点：14:30,15:30,16:30,17:30,18:30,19:30,20:30,21:30,22:30
6 月 27 日	08:10 出发	璟峯酒店→南科大会议中心
	08:10 出发	博林天瑞喜来登酒店→南科大会议中心
	08:10 出发	维也纳好眠国际酒店→南科大会议中心
	18:00 出发	南科大商学院→璟峯酒店
	18:00 出发	南科大商学院→博林天瑞喜来登酒店
	18:00 出发	南科大商学院→维也纳好眠国际酒店

接驳班车

6 月 28 日	09:00 出发	璟峯酒店→南科大会议中心
	09:00 出发	博林天瑞喜来登酒店→南科大会议中心
	09:00 出发	维也纳好眠国际酒店→南科大会议中心
	18:00 出发	南科大商学院→璟峯酒店
	18:00 出发	南科大商学院→博林天瑞喜来登酒店
	18:00 出发	南科大商学院→维也纳好眠国际酒店
	18:00（晚宴去程）	南科大商学院→博林天瑞喜来登酒店→酒店
	21:00（晚宴返程）	
6 月 29 日	07:50 出发	璟峯酒店→南科大会议中心
	08:00 出发	博林天瑞喜来登酒店→南科大会议中心
	08:00 出发	维也纳好眠国际酒店→南科大会议中心
	12:20 出发	南科大商学院→璟峯酒店
	12:20 出发	南科大商学院→博林天瑞喜来登酒店
	12:20 出发	南科大商学院→维也纳好眠国际酒店

Meals

To facilitate informal discussions among participants, complimentary lunch and coffee breaks will be provided. In addition, the conference banquet will be held on June 28 at 18:30. Please note that the banquet is not included in the registration fee. Participants who wish to attend may purchase tickets separately on a voluntary basis. Tickets can be obtained at the registration desk upon inquiry and purchase. The deadline for dinner registration and payment is 17:00 on June 27. We look forward to your participation.

Dining location & Times

Venue: University Center Cafeteria; Sheraton Shenzhen Nanshan

Date	Meal	Time	Location
June 27	Lunch	12:00- 14:00	University Center Cafeteria
June 28	Lunch	12:00- 14:00	University Center Cafeteria
June 28	Banquet (tickets required)	18:30- 21:00	Sheraton Shenzhen Nanshan
June 29	Lunch	12:00- 14:00	University Center Cafeteria

Address:

(1) University Center Cafeteria: The on-campus main dining hall situated on the ground floor of the SUSTech University Center in the central campus. Facing the campus landmark Big Banyan Tree Plaza, it is adjacent to the Yidan Library and the Second Research Building.

(2) Sheraton Shenzhen Nanshan: No. 4088, Liuxian Boulevard, Nanshan District, Shenzhen.

用餐

为方便参会嘉宾在会议期间自由交流、深入探讨，本次大会将提供午餐及茶歇。另设大会晚宴，晚宴不包含在会务注册费内，有意愿参与的嘉宾可自愿单独购票参与，具体可至签到台咨询购买和领取晚宴票。晚宴定于 6 月 28 日 18:30 举办，晚宴报名缴费截止时间为 6 月 27 日 17:00。

日期	用餐	用餐时段	用餐地点
6 月 27 日	午餐	12:00–14:00	南科大中心餐厅
6 月 28 日	午餐	12:00–14:00	南科大中心餐厅
6 月 28 日	晚宴（凭票入场）	18:30–21:00	博林天瑞喜来登酒店
6 月 29 日	午餐	12:00–14:00	南科大中心餐厅

用餐地址：

（1）南科大中心餐厅：校内食堂，位于校园核心南科大中心一楼，门前为校园地标大榕树广场，毗邻一丹图书馆、第二科研楼。

（2）博林天瑞喜来登酒店：深圳市南山区留仙大道 4088 号。

Program Overview

Day 1 (June 27, Saturday, 2026)			
Time	Venue	Agenda Items	Host
09:00AM-09:25AM	Concert Hall, 1/F of Convention Center, SUSTech; Round Table Meeting Room, 4/Fof Convention Center, SUSTech	Welcome and Opening Ceremony	Jianguo Sun
09:25AM-09:30AM		Group Photo	
09:30AM-10:30AM		Keynote Lecture 1-Jeremy Taylor	Jingyuan Liu
10:30AM-11:00AM		Coffee Break	
11:00AM-12:00PM		Keynote Lecture 2-Annie Qu	Qiman Shao
12:00PM-02:00PM	University Center Cafeteria	Lunch	
02:00PM-03:40PM	Chi Wah Lecture Hall(Third Teaching Building)	Parallel Invited Sessions	Session Chairs
03:40PM-04:00PM		Coffee Break	
04:00PM-05:40PM		Parallel Invited Sessions	Session Chairs
Day 2 (June 28, Sunday, 2026)			
09:30AM-10:30AM	Concert Hall, 1/F of Convention Center, SUSTech; Round Table Meeting Room, 4/Fof Convention Center, SUSTech	Keynote Lecture 3-Niansheng Tang	Jianguo Sun
10:30AM-11:00AM		Coffee Break	
11:00AM-12:00PM		Keynote Lecture 4-Weijie Su	Xuejun Jiang
12:00PM-02:00PM	University Center Cafeteria	Lunch	
02:00PM-03:40PM	Chi Wah Lecture Hall(Third Teaching Building)	Parallel Invited Sessions	Session Chairs
03:40PM-04:00PM		Coffee Break	
04:00PM-05:40PM		Parallel Invited Sessions	Session Chairs
06:30PM-09:00PM	Sheraton Shenzhen Nanshan	Conference Banquet	
Day 3 (June 29, Monday, 2026)			
08:30AM-10:10AM	Chi Wah Lecture Hall(Third Teaching Building) Business School Building	Parallel Invited Sessions	Session Chairs
10:10AM-10:30AM		Coffee Break	
10:30AM-12:10PM		Parallel Invited Sessions	Session Chairs
12:10PM-02:00PM	University Center Cafeteria	Lunch	
ICSA China Conference, Shenzhen, Guangdong, China, June 27-June 29, 2026			



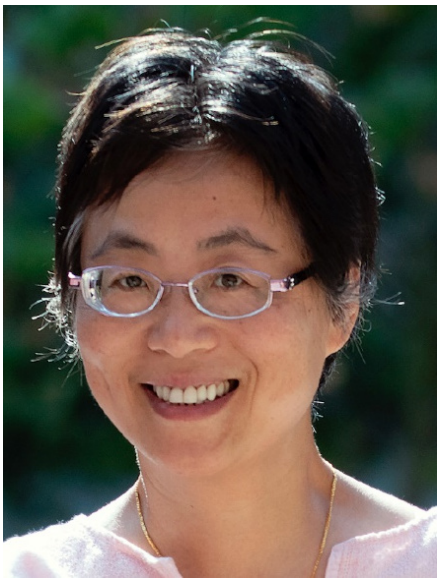
Jeremy Taylor, PhD, is the Pharmacia Research Professor of Biostatistics in the School of Public Health and Professor in the Department of Radiation Oncology in the School of Medicine at the University of Michigan, Ann Arbor. He is the director of the University of Michigan Cancer Center Biostatistics Unit and director of the Cancer/Biostatistics training program. He received his B.A. in Mathematics from Cambridge University and his Ph.D. in Statistics from UC Berkeley. He was on the faculty at UCLA from 1983 to 1998, when he moved to the University of Michigan. He has had visiting positions at the Medical Research Council, Cambridge, England; the University of Adelaide; INSERM, Bordeaux and CSIRO, Sydney, Australia. He is a previously winner of the Mortimer Spiegelman Award from the American Public Health Association and the Michael Fry Award from the Radiation Research Society. He has worked in various areas of Statistics and Biostatistics, including Box-Cox transformations, longitudinal and survival analysis, cure models, missing data, smoothing methods, clinical trial design, surrogate and auxiliary variables. He has been heavily involved in collaborations in the areas of radiation oncology, cancer research and bioinformatics.

Time: 09:30 AM to 10:30 AM, June 27, 2026

Title: Data Integration: General Concepts and Specific Methods

Abstract:

I will consider the situation where you have a small or modest sized dataset and you wish to build a prediction model for the distribution of Y given covariates X and Z, where X and Z are low dimensional, with X being commonly collected and Z being only available in your dataset. You have external summary information from one or more studies about the distribution of Y given a subset of X. The individual data from these external studies is not available to you. The main goal is to improve the statistical efficiency of the model for Y given X and Z by making use of the external summary information. I will discuss empirical likelihood methods, Bayesian approaches, synthetic data methods and estimating equation approaches. Key concepts are the degree of similarity that could be assumed about the covariate distributions and the Y given X distributions, between the internal and external populations. Shrinkage approaches including James-Stein and Kullback-Leibler based shrinkage can be used to give estimators that are robust to population heterogeneity.



Annie Qu is Professor at the Department of Statistics and Applied Probability, University of California, Santa Barbara starting July 2025. She received her Ph.D. in Statistics from the Pennsylvania State University in 1998. Qu's research focuses on solving fundamental issues regarding structured and unstructured large-scale data and developing cutting-edge statistical methods and theory in machine learning and algorithms for personalized medicine, text mining, recommender systems, medical imaging data, and network data analyses for complex heterogeneous data. Dr. Qu was a Data Science Founder Professor of Statistics and the Director of the Illinois Statistics Office at the University of Illinois at Urbana-Champaign during her tenure in 2008-2019, and Chancellor's Professor at UC Irvine in 2020-2025. She was a recipient of the NSF Career award from 2004 to 2009. She is a Fellow of the Institute of Mathematical Statistics (IMS), the American Statistical Association, and the American Association for the Advancement of Science. She is also a recipient of IMS Medallion Award and Lecturer in 2024. She served as Journal of the American Statistical Association Theory and Methods Co-Editor from 2023 to 2025, IMS Program Secretary from 2021 to 2027 and ASA Council of Sections of Governing Board Chair in 2025. She is the recipient of the 2025 Carver Medal of IMS.

Time: 11:00 AM to 12:00 PM, June 27, 2026

Title: Representation Retrieval Learning for Heterogeneous Data Integration

Abstract:

In the era of big data, large-scale, multi-modal datasets are increasingly ubiquitous, offering unprecedented opportunities for predictive modeling and scientific discovery. However, these datasets often exhibit complex heterogeneity, such as covariate shift, posterior drift, and missing modalities which can hinder the accuracy of existing prediction algorithms. To address these challenges, we propose a novel Representation Retrieval (R2) framework, which integrates a representation learning module (the representer) with a sparsity-induced machine learning model (the learner). Moreover, we introduce the notion of "integrativeness" for representers, characterized by the effective data sources used in learning representers, and propose a Selective Integration Penalty (SIP) to explicitly improve the property. Theoretically, we demonstrate that the R2 framework relaxes the conventional full-sharing assumption in multi-task learning, allowing for partially shared structures, and that SIP can improve the convergence rate of the excess risk bound. Extensive simulation studies validate the empirical performance of our framework, and applications to two real-world datasets further confirm its superiority over existing approaches.



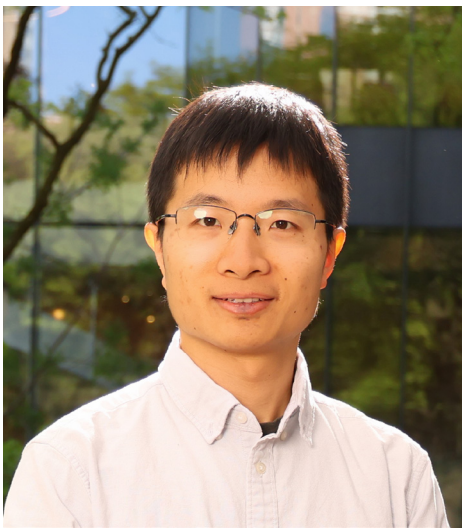
Niansheng Tang（唐年胜）, Professor and Doctoral Supervisor at Yunnan University, recipient of the National Science Fund for Distinguished Young Scholars, Distinguished Professor of the Cheung Kong Scholars Program of the Ministry of Education, National-Level Distinguished Teacher, selected for the National Hundred, Thousand, and Ten Thousand Talents Project, honored as a National "Young and Middle-Aged Expert with Outstanding Contributions," Elected Member of the International Statistical Institute (ISI Elected Member), Fellow of the Institute of Mathematical Statistics (IMS Fellow), member of the Teaching Steering Committee for Statistics and Related Majors in Higher Education Institutions under the Ministry of Education, and member of the 8th and 9th Academic Leadership Group of the Mathematics Tianyuan Fund of the National Natural Science Foundation of China. His research focuses on biostatistics, high-dimensional data analysis, missing data analysis, Bayesian statistics, machine learning, variational inference, and related areas. He has published over 190 academic papers in journals such as JASA, Annals of Statistics, and Biometrika, and has authored 5 academic monographs. He currently serves as an editorial board member or associate editor for journals such as Statistics and Its Interface and Communications in Mathematics and Statistics.

Time: 9:30 AM to 10:30 AM, June 28, 2026

Title: Variational Masking Generative Model for Anomaly Detection on Incomplete Tabular Data

Abstract:

Building anomaly detection models solely from normal data is crucial for many real-world applications. Existing deterministic autoencoder-based methods rely on reconstruction errors for anomaly detection but lack explicit modeling of probability distribution of the normal data. This limitation inherently hinders their performance and applicability when handling incomplete data. This paper proposes a novel Variational Masking Generative Model (VMGM) for tabular data anomaly detection. Specifically, we learn the data distribution via introducing learnable masks and modeling conditional distribution of the masked features of data given the unmasked features using a conditional diffusion model. By treating these masks as latent variables and approximating their intractable posterior with a variational posterior, we derive an evidence lower bound (ELBO) for the model. The learnable masks are sampled using differentiable Gumbel-Sigmoid reparameterization. This approach enables us to derive the parametric form of the ELBO for end-to-end model optimization. For anomaly detection, we construct a detection score based on the discrepancy between generated masked data and the ground-truth masked data. Extensive experiments on six diverse tabular datasets show that VMGM outperforms several state-of-the-art methods. Specifically, it improves AUROC and AUPRC by an average of 1.34% and 1.35%, respectively, over the leading baseline for anomaly detection on both complete and incomplete data.



Weijie Su（苏炜杰）is an Associate Professor in the Wharton Statistics and Data Science Department and, by courtesy, in the Departments of Computer and Information Science, and Biostatistics, Epidemiology and Informatics, at the University of Pennsylvania. He serves as a co-director of the Penn Research in Machine Learning (PRiML) Center. Prior to joining Penn, he earned his PhD in from Stanford University in 2016 and his bachelor's degree from Peking University in 2011. His research interests include statistical foundations of generative AI, privacy-preserving data analysis, high-dimensional statistics, and optimization. He serves as a co-editor of Statistical Learning and Data Science (SLADS) and an associate editor of Journal of the American Statistical Association, Journal of Machine Learning Research, Annals of Applied Statistics, Operations Research, Harvard Data Science Review, and Foundations and Trends in Statistics. He is currently on the ICML 2026 Organizing Committee as the Scientific Integrity Chair.

Time: 11:00 AM to 12:00 PM, June 28, 2026

Title: Alignment in Large Language Models: Statistical and Game-Theoretic Perspectives

Abstract:

Large language models (LLMs) are predominantly aligned with human preferences through reinforcement learning from human feedback (RLHF). In this talk, we explore the theoretical foundations of LLM alignment through the intertwined lenses of statistics and game theory. First, we show how the current formulation of RLHF induces a systematic bias we call preference collapse, and how this can be mitigated by introducing a tailored regularization term into the reward function. Next, we expose a fundamental bottleneck of reward-based alignment, demonstrating that cyclic human preferences cannot be faithfully represented by scalar reward models such as the Bradley-Terry model. More precisely, we establish that such cyclic inconsistencies give rise to a lower bound on the approximation error of any scalar reward fitting. Shifting to a game-theoretic perspective, we focus on Nash learning from human feedback and establish several social choice desiderata for this approach to alignment, including the preservation of preference diversity through the emergence of mixed strategies. Finally, we show that the zero-sum game approach generally cannot perfectly match a target preference distribution as a unique Nash equilibrium.

Junior Researcher Award Winners

Changxiao Cai, University of Michigan
Title: Minimax Optimality of the Probability Flow Ode for Diffusion Models

Xiaoyu Hu, Xi'an Jiaotong University
Title: Neural Wasserstein Two-Sample Tests

Yichen Lou, Nanyang Technological University, Singapore
Title: Transfer Learning Estimation of the Accelerated Failure Time Model Based on High-Dimensional Data

Zhimei Ren, University of Pennsylvania
Title: Boosting e-BH via conditional calibration

Yuxin Tao, Southern University of Science and Technology
Title: Homogeneity Pursuit in Ranking Inference Based on Pairwise Comparison

Yuchen Zhou, University of Illinois Urbana-Champaign
Title: Heteroskedastic Tensor Clustering

(The award ceremony will be held during the banquet)

Poster Session

Poster Session Time: 06:00 PM to 10:00 PM, June 26

Poster Evaluation Time: 06:00 PM to 08:00 PM, June 26

Venue: Outside Dacheng Auditorium (Lecture Hall 114), Business School Building

All poster presenters must be present on-site for presentation during the poster judging session.

(The award ceremony will be held during the banquet)

No.	Name	Affiliation	Title
1	Dan Bai	Caidya	Application of Adaptive Phase II/III Seamless Designs for Dose Selection: A Comparative Study of Inferentially and Operationally Seamless Design
2	Zhiqiang Cao	Shenzhen Technology University	Covariate-Adjusted Win Statistics in Randomized Clinical Trials with Ordinal Outcomes
3	Junzhuo Gao	City University of Hong Kong	Power-Enhancing Probability Subsampling Using Side Information
4	Nazanin Hashemi	The First Affiliated Hospital of Xi'an Jiaotong University	Gut Microbiome Dysbiosis as a Potential Mediator Linking Perfluorooctanoic Acid (PFOA) Exposure to Colorectal Carcinogenesis: A Mechanistic Evidence-Based Review
5	Xin Huang	Towson University	Self-Supervised Learning for Automated Mouse Brain Extraction in Data-Constrained MRI Studies
6	Ying Huang	Fred Hutchinson Cancer Center	Assessing Screening Efficacy in the Presence of Cancer Overdiagnosis
7	Wenxin Jiang	City University of Hong Kong	TraceCB: Trans-Ancestry Cell-Type-Specific eQTLs Mapping by Integrating scRNA-Seq and Bulk Data
8	Kai Kang	University of North Carolina Wilmington	A Hierarchical Spatial Fractional Binomial Model for Spatial Transcriptomics and Cross-Sample Comparison
9	Eunjung Kim	Konkuk University, Korea	Doubly Robust Estimators for Treatment-Specific Cumulative Incidence Functions under Competing Risks
10	Jian Lei	Beijing Normal University	Statistical Inference for High-Dimensional Logistic Regression with Latent Random Change Point

11	Hanwen Liu	Southeast University	Corrected Convolution-Type Loss for Quantile Regression Estimation with Laplace-Distributed Covariate Measurement Errors
12	Liangbo Lyu	University of Michigan	Leveraging Text Covariates to Improve Conformal Causal Inference on Randomized Trials
13	Xi Ning	Colby College	Regression Analysis of Semiparametric Cox-Aalen Transformation Models with Partly Interval-Censored Data
14	Jinglan Qiu	University of Victoria	Spatial Bayesian Framework for Mapping Long Memory in Resting-State fMRI
15	Seungjun Ryu	Konkuk University	Neural Network Cause-Specific Cox Modeling for Right-Censored Competing Risks
16	Dong Yeon Shin	Konkuk University	PI Construction with Random Forests for Longitudinal Data
17	Jie Tang	Beijing Normal University	Empirical Likelihood-Based Fairness Auditing: Distribution-Free Certification and Flagging
18	Minghan Wang	Beijing Normal University	Pursuing Goodness-of-Fit of Quantile Regressions via Random Projections
19	Jiazhen Xu	Macquarie University	Spherically Embedded Time Series with Unknown Trend and Periodic Components
20	Tingting Zhao	University of Rhode Island	Safe and Swift: A Game-Theoretic Framework for Rapid Detection of Volatility Regime Shifts in Financial Markets
21	Xinkai Zhou	Beijing Normal-Hong Kong Baptist University	FLAME: A Model for Duration-Dependent Risk Accumulation in Episodic Temporal Exposures

Short Course

Course Title: Statistical and Algorithmic Foundations of Diffusion Models

Time: 6/26 Friday 8:30AM-12:30PM (Half Day)

Venue: Room 104, Business School Building, SUSTech

Abstract:

Diffusion generative models have emerged as a cornerstone of modern generative AI, delivering state-of-the-art performance across a wide range of data generation tasks. At their core, diffusion models seek to gradually transform pure noise into new data samples that emulate a target data distribution, accomplished by learning to reverse a forward stochastic process that progressively converts data into Gaussian noise. Despite their empirical successes, the statistical and algorithmic foundations of diffusion models remain far from mature. This lack of fundamental understanding limits their broader adoption, especially in applications that demand interpretability and reproducibility.

This short course provides a timely introduction to diffusion models and presents recent progress toward understanding their striking effectiveness, with an emphasis on core principles and statistical insights. We will examine the fundamental mechanisms of score-based diffusion models; characterize the statistical limits of learning score functions; analyze the convergence behavior of diffusion-based samplers; explore how these models adapt to unknown low-dimensional data structures; discuss conditional generation via diffusion guidance; and highlight ideas for accelerating inference through higher-order approximations. Throughout this short course, we will connect theoretical advances to practical applications, illustrating how fundamental insights can inform effective algorithm design.

Lecturer 1: Yuxin Chen

Professor, Department of Statistics and Data Science

Wharton School at the University of Pennsylvania

Lecturer 2: Gen Li

Assistant Professor, Department of Statistics

Chinese University of Hong Kong

Lecturer 3: Yuting Wei

Associate Professor, Department of Statistics and Data Science

Wharton School at the University of Pennsylvania

Parallel Invited Sessions Schedule

a. June 27th PM (14:00-15:40)

Number	Title	Organizer	Chair	Room
S007	Advanced Statistical Modeling and Integration of Complex Data	Fengqing Zhang	Fengqing Zhang	Chi Wah Lecture Hall 103
S010	Advances in High-Dimensional and Nonparametric Statistics	Tao Hu	Yukun Liu	Chi Wah Lecture Hall 105
S013	Advances in Modeling Frameworks for Structured and Complex Data	Yan Lu	Yan Lu	Chi Wah Lecture Hall 201
S023	Advances in Time Series Modeling and Functional Data Analysis	Lizhe Sun	Lizhe Sun	Chi Wah Lecture Hall 302
S028	Bayesian Regression and Variable Selection for Structured and High-Dimensional Data	Xia Wang	Xia Wang	Chi Wah Lecture Hall 102
S031	Bridging Statistics and Modern AI: Foundations for Deep Learning and Generative Models	Yuxin Chen	Yuxin Chen	Chi Wah Lecture Hall 104
S033	Computational and Statistical Advances in Association Modeling	Shili Lin	Shili Lin	Chi Wah Lecture Hall 106
S036	Data Integration UnderDifferent Data Types: New Methodology Developments	Peisong Han	Yukun Liu	Chi Wah Lecture Hall 112
S038	Emerging Statistical Frameworks for High-Dimensional and Biased Data	Shouhao Zhou	Shouhao Zhou	Chi Wah Lecture Hall 107
S039	Estimation and Testing in Sparse, Quantile, and Survival Models	Gang Li (Co-organized with Hong Wang)	Hong Wang	Chi Wah Lecture Hall 203
S042	From Cells to Tissues: Statistical Frameworks for Spatial Omics	Xiang Zhou	Xiang Zhou	Chi Wah Lecture Hall 202
S055	Innovations in Regression and Change Detection for Complex Data	Xinyuan Song	Xiangnan Feng	Chi Wah Lecture Hall 108
S073	Modern Statistical Methods for Complex Data: Hypothesis Testing, Estimation and Application	Bingshu Chen	Bingshu Chen	Chi Wah Lecture Hall 301
S088	Precision Medicine and High-Dimensional Data Inference	Xuerong Chen	Xuerong Chen	Chi Wah Lecture Hall 204
S097	Recent Advances in Data-Driven Decision Making	Lu Wang	Lu Wang	Chi Wah Lecture Hall 205
S104	Recent Advances in Omics Data Analysis Techniques	Tianwei Yu	Tianwei Yu	Chi Wah Lecture Hall 303
S109	Recent Advances in Statistical Designs for Early-Phase Oncology Clinical Trials	Yong Zang	Yan Han	Chi Wah Lecture Hall 304
S114	Recent Advances on the Analysis of Complex Data	Jianguo Sun	Jianguo Sun	Chi Wah Lecture Hall 210
S126	Statistical and Computational Methods for Omics and Biological Discovery	Liangliang Zhang	Liangliang Zhang	Chi Wah Lecture Hall 305

S128	Statistical Foundations for Machine Learning and AI	Weijie Su	Xiang Li	Chi Wah Lecture Hall 206
S050	Frontiers of Transfer Learning, Data Fusion, and Adaptive Evidence Generation: Theory, Methods, and Practice	Kun Chen	Jianmin Chen	Chi Wah Lecture Hall 207
S136	Statistical Learning and Modeling for Complex Data	Zheyang Wu	Donghui Yan	Chi Wah Lecture Hall 306
S147	Study Design and Joint Modeling with Longitudinal and Survival Data	Zhigang Li	Yang Yang	Chi Wah Lecture Hall 307
S154	When Large Language Model Meets Statistics	Jingyuan Liu	Jingyuan Liu	Chi Wah Lecture Hall 208
S005	Advanced Statistical Methods for Complex Survival Data	Peijie Wang	Peijie Wang	Chi Wah Lecture Hall 308

b. June 27th PM (16:00-17:40)

Number	Title	Organizer	Chair	Room
S021	Advances in Statistical Modeling for Causal Inference and High-dimensional Complex Data	Peijie Wang	Peijie Wang	Chi Wah Lecture Hall 304
S025	AI and Big Data Analysis	Yichuan Zhao	Ting Zhang	Chi Wah Lecture Hall 102
S030	Bootstrap and Model Averaging Methods for Robust Statistical Learning	Chenglong Ye	Xiaofang Shi	Chi Wah Lecture Hall 208
S041	Advances in Statistical Inference for Dependent and High-Dimensional Data	Xuejun Jiang	Jiancheng Jiang	Chi Wah Lecture Hall 103
S045	Frontiers in Biostatistics: Robust Inference and Machine Learning	Jialiang Li	Mu Yue	Chi Wah Lecture Hall 104
S129	Statistical Foundations for Preference Learning, Representation, and Heterogeneity in the Era of AI	Yong Chen	Jianmin Chen	Chi Wah Lecture Hall 106
S054	Inference and Learning Under Heterogeneous Data	Nan Lin	Nan Lin	Chi Wah Lecture Hall 305
S056	Innovative Approaches in Machine Learning and Health Data Analytic	Chen-An Tsai	Chung Chang	Chi Wah Lecture Hall 107
S058	Innovative Statistical Design Methods for Modern Clinical Trials	Shouhao Zhou	Yuexi Wang	Chi Wah Lecture Hall 108
S059	Integrative Statistical Methods for Genetic Association and Multi-Omics Studies	Xiang Zhou	Xiang Zhou	Chi Wah Lecture Hall 202
S075	Modern Statistical Methods for High-Dimensional and Structured Data	Xinyuan Song	Kai Kang	Chi Wah Lecture Hall 105
S076	Modern Statistical Methods for High-Dimensional Data and Machine Learning Models	Quefeng Li	Zhao Chen	Chi Wah Lecture Hall 112
S079	New Advances in Infectious Disease Modeling	Zhigang Li	Zhigang Li	Chi Wah Lecture Hall 203

S107	Recent Advances in Spatial Omics and Single-cell Data Analysis	Yingying Wei	Fangda Song	Chi Wah Lecture Hall 306
S112	Recent Advances of Time Series Analysis and Change-Point Detection	Zhanrui Cai	Yuxin Tao	Chi Wah Lecture Hall 201
S117	Recent Developments on Applied Statistics	Tsung-I Lin	Chang-Yun Lin	Chi Wah Lecture Hall 301
S122	Some New Methods of Robust Statistical Learning for Complex Data	Hui Zhao	Xiaofei Zhang	Chi Wah Lecture Hall 204
S124	Statistical Analysis of Complex Data	Xinyu Zhang	Ziwen Gao	Chi Wah Lecture Hall 205
S127	Statistical and Computational Methods for Single-Cell, Spatial, and Multi-Omics Integration	Hongkai Ji	Hongkai Ji	Chi Wah Lecture Hall 210
S130	Statistical Foundations of LLM Behavior, Alignment, and Automated Reasoning	Chenglong Ye	Chenglong Ye	Chi Wah Lecture Hall 206
S131	Statistical Frontiers in Spatial Omics	Jingyi Jessica Li	Guan' ao Yan	Chi Wah Lecture Hall 207
S140	Statistical Methods and Applications	Jianguo Sun	Jianguo Sun	Chi Wah Lecture Hall 302
S148	Survival Analysis and Related Statistical Methods	Tao Hu	Yijun Wang	Chi Wah Lecture Hall 307
S106	Recent Advances in Sequential Methodologies	Yiming Xing	Ruizhi Zhang	Chi Wah Lecture Hall 308
S012	Advances in Machine Learning Methods for Survival and Event-Time Data	Gang Li	Wen Su	Chi Wah Lecture Hall 303

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

Number	Title	Organizer	Chair	Room
S002	Advanced Methods for Complex Data Analysis	Po Yang	Po Yang	Chi Wah Lecture Hall 302
S092	Recent Advances for the Analysis of Complex Data	Mingyue Du	Mingyue Du	Chi Wah Lecture Hall 102
S014	Advances in Modern High-dimensional Statistics	Anru Zhang	Haoxuan Sun	Chi Wah Lecture Hall 103
S022	Advances in Survival Analysis and Causal Inference	Xingqiu Zhao	Xingqiu Zhao	Chi Wah Lecture Hall 105
S019	Advances in Statistical Methods for Clinical Trials: Longitudinal Data and Complex Data Mechanisms	Wei Tu	Wei Tu	Chi Wah Lecture Hall 303
S029	Big Data and AI	Mei-Ling Ting Lee	Chunling Liu	Chi Wah Lecture Hall 104
S040	Experimental Design and Optimization	Fasheng Sun	Fasheng Sun	Chi Wah Lecture Hall 304

S052	High-Dimensional Inference and P-Value Combination	Qizhai Li	Haili Zhang	Chi Wah Lecture Hall 112
S062	Machine Learning and Statistical Modeling Applications	Fasheng Sun	Fasheng Sun	Chi Wah Lecture Hall 106
S063	Microbiome Intelligence: Large Models, Big Data, and Statistical Advances	Liangliang Zhang	Ni Zhao	Chi Wah Lecture Hall 202
S069	Modern Statistical Inference and Modeling for Complex and Structured Data	Fengting Yi	Fengting Yi	Chi Wah Lecture Hall 203
S070	Modern Statistical Learning and High Dimensional Modeling	Xingqiu Zhao	Xingqiu Zhao	Chi Wah Lecture Hall 204
S078	New Advances in Differential Privacy	Xuerong Chen	Yanqing Zhang	Chi Wah Lecture Hall 205
S082	New Developments of Statistical Methods in Bio-Medical Studies	Gengsheng Qin	Gengsheng Qin	Chi Wah Lecture Hall 305
S086	Novel Statistical and Machine Learning Methods for Analyzing Omics Data	Shili Lin	Hongmei Jiang	Chi Wah Lecture Hall 107
S099	Recent Advances in High-Dimensional and Functional Hypothesis Testing	Fengting Yi	Fengting Yi	Chi Wah Lecture Hall 210
S100	Recent Advances in High-Dimensional Data Analysis	Chiung-Yu Huang	Ying Sheng	Chi Wah Lecture Hall 306
S115	Recent Advances on the Analysis of Failure Time Data	Jianguo Sun	Mingyue Du	Chi Wah Lecture Hall 307
S121	Scalable and Distributed Statistical Learning for Complex Data Structures	Lu Tang	Lu Tang	Chi Wah Lecture Hall 108
S135	Statistical Innovations for Spatial Transcriptomics and 3D Genomics	Li-Xuan Qin	Jian Zou	Chi Wah Lecture Hall 201
S139	Statistical Learning with AI	Jiancheng Jiang	Jiancheng Jiang	Chi Wah Lecture Hall 206
S146	Structured Data Analysis and Dependence Modeling	Fang Yao	Ying Yang	Chi Wah Lecture Hall 207
S150	The Frontier Research of Statistical Process Quality Control Under AI Era	Chunjie Wu	Chunjie Wu	Chi Wah Lecture Hall 301
S144	Statistics and Machine Learning at the Frontiers of Genomics and Genetics	Hongkai Ji	Hongkai Ji	Chi Wah Lecture Hall 208

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

Number	Title	Organizer	Chair	Room
S001	Advance in Statistical Methods for Genetic and Genomic Studies	Jin Liu	Jin Liu	Chi Wah Lecture Hall 201
S003	Advanced Methods for Dependence, Signals and Spatio-Temporal Detection	Zheyang Wu	Wen Zhou	Chi Wah Lecture Hall 202

S015	Advances in Predictive and Integrative Modeling for Complex Health Data	Ming Wang	Lijun Zhang	Chi Wah Lecture Hall 211
S016	Advances in Statistical Inference for Complex and High-Dimensional Models	Fang Yao	Fang Yao	Chi Wah Lecture Hall 102
S026	AI-Driven Genomics and Precision Medicine: From Prediction to Personalization	Ching-Ti Liu	Xiaoyu Song	Chi Wah Lecture Hall 104
S024	Advancing Complex Dependent Data Analysis	Fukang Zhu	Zhengwei Liu	Chi Wah Lecture Hall 301
S034	Conditional Modeling and Causality: Generative Models for Learning and Integration	Jian Huang	Jian Huang	Chi Wah Lecture Hall 106
S047	Frontiers in Discrete-Valued Time Series Modeling	Fukang Zhu	Fukang Zhu	Chi Wah Lecture Hall 303
S048	Frontiers in Learning and Optimization: From GenAI to Dynamic Operations	Zhengling Qi	Zeyu Bian	Chi Wah Lecture Hall 107
S049	Frontiers in Methods for Complex-structured Survival Data	Huiqiong Li	Huiqiong Li	Chi Wah Lecture Hall 103
S060	Interval-Valued Data Modelling and Forecasting	Yuying Sun	Yuying Sun	Chi Wah Lecture Hall 105
S066	Modern Methods for Omics Inference and Clinical Translation	Ming Wang	Weihong Guo	Chi Wah Lecture Hall 304
S068	Modern Perspectives on High-Dimensional Statistical Theory and Methodology	Tianying Wang	Yuqian Zhang	Chi Wah Lecture Hall 305
S071	Modern Statistical Learning for Complex Data Analysis	Lexin Li	Ting Li	Chi Wah Lecture Hall 112
S084	New Methods for the Analysis of Complex Data	Mingyue Du	Mingyue Du	Chi Wah Lecture Hall 203
S085	New Statistical and Coputational Methods for Large Data Analysis	Tingting Zhang	Ziqi Chen	Chi Wah Lecture Hall 204
S096	Recent Advances in Computational Methods for Single-Cell and Spatial Omics	Ruibin Xi	Ruibin Xi	Chi Wah Lecture Hall 205
S103	Recent Advances in Network Analysis	Wei Lan	Wei Lan	Chi Wah Lecture Hall 210
S118	Recent Progress in Dependent Data Analysis	Tsung-I Lin	Wei-Ying Wu	Chi Wah Lecture Hall 306
S133	Statistical Inference for Complex Data	Jingyuan Liu	Jingyuan Liu	Chi Wah Lecture Hall 108
S141	Statistical Methods for Genetics and Genomics	Zheyang Wu	Zheyang Wu	Chi Wah Lecture Hall 308
S142	Statistical Methods for Reliable Machine Learning	Hongxin Wei	Hongxin Wei	Chi Wah Lecture Hall 206

S143	Statistical Modeling and Inference for Large Language Models	Weijie Su	Chendi Wang	Chi Wah Lecture Hall 207
S149	The Broad Roles of Design in Modern Statistical Analysis	Boxiang Wang	Lyuou Zhang	Chi Wah Lecture Hall 302
S153	Uncertainty Quantification in Modern Statistics and Machine Learning	Chiung-Yu Huang	Huaqing Jin	Chi Wah Lecture Hall 208
S155	Matrix Data	Wang Zhou	Jiang Hu	Chi Wah Lecture Hall 309
S032	Causal Learning in Modern Health and Survey Data	Lu Tang	Lu Tang	Chi Wah Lecture Hall 307

e. June 29th AM (08:30-10:10)

Number	Title	Organizer	Chair	Room
S011	Advances in High-Dimensional Dependence Modeling	Ching-Kang Ing	Ching-Kang Ing	Chi Wah Lecture Hall 115
S018	Advances in Statistical Machine Learning Methods for Complex Data	Changcheng Li	Changcheng Li	Chi Wah Lecture Hall 106
S037	Decoding Complexity: Genomic Diversity, Disease Mechanisms, and Evolutionary Insights	Ching-Ti Liu	Zuoheng Wang	Chi Wah Lecture Hall 204
S043	From Data to Biomarkers: Statistical Innovation for Complex Challenges in Modern Genetics, Genomics and Proteomics	Li-Xuan Qin	Li-Xuan Qin	Chi Wah Lecture Hall 107
S044	From Probabilistic Theory to Causal Inference in Complex Systems	Doudou Zhou	Shuting Shen	Chi Wah Lecture Hall 205
S051	High Dimensional Data Analysis and Change Point Inference	Long Feng	Long Feng	Chi Wah Lecture Hall 201
S064	Model Averaging and Related Topics	Xinyu Zhang	Huihang Liu	Chi Wah Lecture Hall 108
S065	Model Averaging for Time Series Data	Yuying Sun	Yuying Sun	Chi Wah Lecture Hall 210
S072	Modern Statistical Methods and Artificial Intelligence: From Theory to Biomedical Applications	Ling Zhou	Ling Zhou	Chi Wah Lecture Hall 302
S077	Mordern Distributional Methods for Learning, Inference, and Robustness	Jian Huang	Jian Huang	Chi Wah Lecture Hall 202
S081	New Developments in Survey Sampling	Guohua Zou	Wangxue Chen	Chi Wah Lecture Hall 307
S083	New Frontiers in Spatial Modeling: Dissimilarity, Extremes, and High-Dimensional Clustering	Kun Chen	Likun Zhang	Chi Wah Lecture Hall 206

S087	Post and Semi-Supervised Inference	Qizhai Li	Haojie Ren	Chi Wah Lecture Hall 211
S090	Recent Advance in Statistical Learning	Biao Cai	Biao Cai	Chi Wah Lecture Hall 212
S091	Recent Advances for Statistical Analysis of Censored Data	Hui Zhao	Hui Zhao	Chi Wah Lecture Hall 303
S093	Recent Advances in Complex Data Analysis and New Applications	Yichuan Zhao	Lei Huang	Chi Wah Lecture Hall 308
S098	Recent Advances in Designs and Statistical Methods for Clinical Trials Research	Yuan Wu	Xiaofei Wang	Chi Wah Lecture Hall 309
S105	Recent Advances in Selective Inference, Data Assimilation, and Functional Data	Anru Zhang	Wanjie Wang	Chi Wah Lecture Hall 304
S111	Recent Advances in Statistical Methods for Genetic Data	Yingying Wei	Yingying Wei	Chi Wah Lecture Hall 207
S119	Regression Modeling and Efficient Estimation for Complex Data	Liuquan Sun	Meiling Hao	Chi Wah Lecture Hall 305
S132	Statistical Inference and Subgroup Analysis for Specialized Data Types	Liuquan Sun	Ye Fan	Chi Wah Lecture Hall 306
S137	Statistical Learning for Complex Data	Huiqiong Li	Huiqiong Li	Chi Wah Lecture Hall 310
S145	Statistics, Machine Learning, and AI	Lexin Li	Ben Dai	Chi Wah Lecture Hall 208
S057	Innovative Machine Learning and Statistical Methods for Complex Biomedical Data	Yuanjia Wang	Kai Kang	Chi Wah Lecture Hall 203
S151	Time Series Econometrics and Applications	Jiancheng Jiang	Shaoxin Hong	Chi Wah Lecture Hall 301

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

Number	Title	Organizer	Chair	Room
S004	Advanced Statistical and Machine Learning Methods in Genetics and Genomics	Tianying Wang	Bo Yu	Chi Wah Lecture Hall 106
S008	Advances in Bayesian Modeling for Big and Complex Data	Tianzhou Ma	Hongjie Ke	Chi Wah Lecture Hall 204
S009	Advances in High-Dimensional and Complex Data Analysis	Changcheng Li	Changcheng Li	Chi Wah Lecture Hall 205
S017	Advances in Statistical Learning and AI for Biological and Health Sciences	Jun Li	Jun Li	Chi Wah Lecture Hall 107
S020	Advances in Statistical Methods for Omics and Health Data Analysis	Tianzhou Ma	Menglu Liang	Chi Wah Lecture Hall 108
S027	Applications in Biomedical Sciences	Grace S. Shieh	Hao Ho	Chi Wah Lecture Hall 305

S046	Frontiers in Causal Inference: Addressing Time-Dependence, Fairness, and Long-Term Effects in Complex Data	Lu Wang	Lu Wang	Chi Wah Lecture Hall 210
S053	High-Dimensional Statistics	Shurong Zheng	Xiaoyi Wang	Chi Wah Lecture Hall 202
S067	Modern Mixture Model Methodology: Advances in Modeling, Inference, and Applications	Pengfei Li	Chunlin Wang	Chi Wah Lecture Hall 306
S080	New Advances in Microbiome Data Analysis and Related Fields	Zhigang Li	Lihui Zhao	Chi Wah Lecture Hall 302
S089	Random Matrix Theory and Applications	Shurong Zheng	Shurong Zheng	Chi Wah Lecture Hall 206
S006	Advanced Statistical Mthods for Network Data and Latent Variable Models	Haoran Zhang	Siliang Zhang	Chi Wah Lecture Hall 115
S094	Recent Advances in AI for Biological Sciences	Ruibin Xi	Ruibin Xi	Chi Wah Lecture Hall 207
S095	Recent Advances in Complex Data Analysis	Chen-Hung Kao	Chen-Hung Kao	Chi Wah Lecture Hall 201
S101	Recent Advances in Longitudinal Data Analysis	Mengying You	Mengying You	Chi Wah Lecture Hall 211
S102	Recent Advances in Modeling Survival Data	Yingwei Peng	Yingwei Peng	Chi Wah Lecture Hall 307
S108	Recent Advances in Spatio-Temporal Data Analysis	Ching-Kang Ing	Shu-Hui Yu	Chi Wah Lecture Hall 212
S110	Recent Advances in Statistical Learning and Big Data	Xiaochao Xia	Xiaochao Xia	Chi Wah Lecture Hall 301
S113	Recent Advances on Statistical Inference and Econometrics	Quefeng Li	Ruoqing Zhu	Chi Wah Lecture Hall 303
S116	Recent Advances in Complex Time Series and Dynamic Network Analysis	Han Yan	Han Yan	Chi Wah Lecture Hall 208
S120	Robust and Flexible Statistical Methods for Complex and High-Dimensional Data	Xuejun Jiang	Hao Lei Weng	Chi Wah Lecture Hall 304
S125	Statistical and Computational Advances for Single-Cell Genomics	Xinzhou Ge	Xinzhou Ge	Chi Wah Lecture Hall 203
S134	Statistical Inference for Complex Data Structures	Chiung-Yu Huang	Fan Xia	Chi Wah Lecture Hall 308
S152	Toward Efficient Clinical Trials: Statistical Methods That Leverage Baseline Covariates	Dong Xi	Zhiwei Zhang	Chi Wah Lecture Hall 309
S074	Modern Statistical Methods for Complex Longitudinal Data and Causal Inference	Xuejun Jiang	Xuejun Jiang	Chi Wah Lecture Hall 310

Detailed Invited Sessions Schedule

a. June 27th PM (14:00-15:40)

S007: Advanced Statistical Modeling and Integration of Complex Data

Room: Chi Wah Lecture Hall 103
Organizer: Fengqing Zhang
Chair: Fengqing Zhang

14:00 Graphical Approaches for Multiplicity Adjustment in Clinical Trials
Jiangtao Gou
Villanova University

14:25 Dynamic Representational Modeling of Language Learning and Processing
Gangyi Feng
The Chinese University of Hong Kong

14:50 How Does Stress Shape Creativity? Evidence from Multimodal Data Analysis of Physiology, Cognition and Neuroscience
• *Haijun Duan, Yifan Wang*
Shaanxi Normal University, Shaanxi Normal University

15:15 A Directed Weighted Graph Neural Network with Tensor Fusion for Multi-Omics Cancer Classification
Tiantian Yang
University of Idaho

S010: Advances in High-Dimensional and Nonparametric Statistics

Room: Chi Wah Lecture Hall 105
Organizer: Tao Hu
Chair: Yukun Liu

14:00 Conformal Prediction Beyond the Horizon: Distribution-Free Inference for Policy Evaluation
Feichen Gan, Youcun Lu, Yingying Zhang, • *Yukun Liu*
East China Normal University, East China Normal University, East China Normal University, East China Normal University

14:25 Estimation and Inference for Density-convoluted Support Vector Machine with Streaming Data
Haochen Rao, • *Xu Guo, Heng Lian, Haobo Qi*
Beijing Normal University, Beijing Normal University, City University of Hong Kong, Beijing Normal University

14:50 Rank Test of Independence Based on Adaptive Hellinger Distance
• *Wenwen Guo, Shilin Shang, Jiujing Wu*
Capital Normal University, Capital Normal University, Capital Normal University

15:15 Linear Spline Index Regression Model: Interpretability, Nonlinearity and Dimension Reduction
Lianqiang Qu, • *Long Lv, Meiling Hao, Liuquan Sun*
Central China Normal University, Central China Normal University, University of International Business and Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences

S013: Advances in Modeling Frameworks for Structured and Complex Data

Room: Chi Wah Lecture Hall 201
Organizer: Yan Lu
Chair: Yan Lu

14:00 A Goodness-of-Fit Test for Zero-Inflated Poisson Mixed Effects Models
• *Juxin Liu, Yanyuan Ma, Jill Johnstone*
University of Saskatchewan, Pennsylvania State University, University of Saskatchewan

14:25 Combined Estimator Using Random Forests and Support Vector Machine in Regression
• *Yan Lu, Ning Xie*
University of New Mexico, Michigan State University

14:50 A Semi-Parametric Model for Small Area Estimation Using Support Vector Machine
• *Guoyi Zhang, John Pleis*
University of New Mexico, National Center for Health Statistics

S023: Advances in Time Series Modeling and Functional Data Analysis

Room: Chi Wah Lecture Hall 302
Organizer: Lizhe Sun
Chair: Lizhe Sun

14:00 Testing and Estimation of Change Point in ARMA Model With Heavy-Tailed G-GARCH Noises
• *Qiang Bai, Shiqing Ling, Ke Zhu*
Shanxi University of Finance and Economics, Hong Kong University of Science and Technology, The University of Hong Kong

14:25 A Class of Adaptive MIDAS Linear Regression Models
Xingfa Zhang, Mengli Liu, Keke Hu
Guangzhou University, Guangzhou University, Guangzhou University

14:50 Variable Selection and Estimation for Generalized Linear Models with Longitudinal Predictors
Xiaojuan Han, Lizhe Sun, Xiaoqin Zhang, Xiang Zhan
Shanxi University of Finance and Economics, Shanxi University of Finance and Economics, Shanxi University of Finance and Economics, Southeast University

15:15 Wavelet-Factor Matrix Hybrid Transport Maps for Spatio-temporal Bayesian Fusion
Shiyuan He
Beijing Technology and Business University

S028: Bayesian Regression and Variable Selection for Structured and High-Dimensional Data

Room: Chi Wah Lecture Hall 102
Organizer: Xia Wang
Chair: Xia Wang

14:00 Integrative Bayesian Modeling for Functional MRI Data
Xuan Cao
University of Cincinnati

14:25 Bayesian Region Selection and Prediction in Poisson Regression with Spatially Dependent Global-Local Shrinkage Prior
Zihan Zhu, Xueying Tang, Shuang Zhou
Yale University, University of Arizona, Arizona State University

14:50 BayesBadgeR: Bayesian Double Generalized Beta Regression with Spatial and Temporal Borrowing for Bounded Well-Being Indices
Abhi Jain,Michael LaValley,Kimberly A. Dukes, Shariq Mohammed
Boston University, Boston University, Boston University,-Boston University

S031: Bridging Statistics and Modern AI: Foundations for Deep Learning and Generative Models

Room: Chi Wah Lecture Hall 104
Organizer: Yuxin Chen
Chair: Yuxin Chen

14:00 Statistical Inference in an Interactive Learning Paradigm
Yuchen Wu, Kangjie Zhou, Weijie Su
Cornell University, Columbia University, University of Pennsylvania

14:25 Transformers Meet In-Context Learning: A Universal Approximation Theory
Gen Li, Yuchen Jiao, Yuting Wei, Yuxin Chen
Chinese University of Hong Kong, Chinese University of Hong Kong, University of Pennsylvania, University of Pennsylvania

14:50 Generative Adversarial Learning with Optimal Input Dimension and Its Adaptive Generator Architecture
Zhiyao Tan, Ling Zhou, Huazhen Lin
Southwestern University of Finance and Economics , Southwestern University of Finance and Economics , Southwestern University of Finance and Economics

15:15 Deep Discrete Encoders: Identifiable Deep Generative Models for Rich Data with Discrete Latent Layers
Seunghyun Lee, Yuqi Gu
Columbia University, Columbia University

S033: Computational and Statistical Advances in Association Modeling

Room: Chi Wah Lecture Hall 106
Organizer: Shili Lin
Chair: Shili Lin

14:00 Adjusting for Non-Confounding Covariates in Case-Control Association Studies
Siliang Zhang, Jinbo Chen, Zhiliang Ying, ◆ Hong Zhang
East China Normal University, University of Pennsylvania,
Columbia University, University of Science and Technology of China

14:25 Integrating Network Annotation From Multiple Correlated Traits To Improve Polygenic Risk Scores Based on GWAS Summary Statistics
Lirong Zhu, Xuewei Cao, Shuanglin Zhang, Qiuying Sha
Tianjin Medical University, Columbia University, Michigan Technological University, Michigan Technological University

14:50 Sparse CCA-Based Mediation Analysis with High-Dimensional Exposures and Mediators
Xincheng Li, Hongmei Jiang
Northwestern University, Northwestern Univeresity

15:15 Cluster Analysis of Longitudinal Associations Between Microbial Abundances and Environmental Factors
Emily Palmer, Caihong Qin, Lan Xue, Yuan Jiang
Hewlett-Packard, Indiana University, Oregon State University, Oregon State University

S036: Data Integration Under Different Data Types: New Methodology Developments

Room: Chi Wah Lecture Hall 112

Organizer: Peisong Han
Chair: Yukun Liu

14:00 A Semiparametric Two-Sample Homogeneity Test with Nonignorable Nonresponse Using Callback Data
Xinyu Wang, Tao Yu, ⋆ *Chunlin Wang, Pengfei Li*
Xiamen University, National University of Singapore, Xiamen University, University of Waterloo

14:25 Orthogonal Covariate Balancing for Causal Inference
Ying Yan
Sun Yat-sen University

14:50 Bias-Corrected Byzantine-Robust Estimator via Cornish-Fisher Expansion for Distributed Learning
Zhonglei Wang
Xiamen University

15:15 Specification Testing in Complex Survey Sampling With Missing Data
⋆ *Puying Zhao, Mengmeng Xu*
Yunnan University, Yunnan University

S038: Emerging Statistical Frameworks for High-Dimensional and Biased Data

Room: Chi Wah Lecture Hall 107
Organizer: Shouhao Zhou
Chair: Shouhao Zhou

14:00 Simulation-Based Inference via Structured Score Matching-Matching
Haoyu Jiang, ⋆ *Yuexi Wang, Yun Yang*
University of Illinois Urbana-Champaign, University of Illinois Urbana-Champaign, University of Maryland, College Park

14:25 High-Dimensional Error-in-Variables Survival Analysis
Jiarui Zhang, Hongsheng Liu, Xin Chen, ⋆ *Jinfeng Xu*
South China of Technology, The University of North Carolina at Chapel Hill, South China University of Science and Technology, City University of Hong Kong

14:50 Addressing Confounding in Comparative Effectiveness Research Using Observational Multilevel Semi-Competing Risks Data
Hyunjae Cho, ⋆ *Hong Zhu*
University of Virginia, University of Virginia

15:15 Capturing High-Dimensional Brain Atrophy Patterns via MRI Distance Geometry and Cross-Attention
Zheyu Wang
Johns Hopkins University

S039: Estimation and Testing in Sparse, Quantile, and Survival Models

Room: Chi Wah Lecture Hall 203
Organizer: Gang Li (Co-organized with Hong Wang)
Chair: Hong Wang

14:00 Orthogonalized Kernel Debiased Machine Learning for Logistic Multimodal Inference
Yaohua Rong
Beijing University of Technology

14:25 Distributed Hard Screening for Quantile Regression
Xianhui Liu
Jiangxi University of Finance and Economics

14:50 SID: A Novel Class of Nonparametric Tests of Independence for Censored Outcomes
Jinhong Li, ⋆ *Jicai Liu, Jinhong You, Riquan Zhang*
Zhejiang Gongshang University, Shanghai Lixin University of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai University of International Business and Economics

S042: From Cells to Tissues: Statistical Frameworks for Spatial Omics

Room: Chi Wah Lecture Hall 202
Organizer: Xiang Zhou
Chair: Xiang Zhou

14:00 Digital Tissue Twins: Uniting Spatial Omics with Histopathology
Mingyao Li
University of Pennsylvania

14:25 Decoupling Biological Signal from Technical Confounding in Library Size for Single Cell and Spatial Omics
Kedi Le, Weiqiang Zhou, ⋆ *Hongkai Ji*
Johns Hopkins Bloomberg School of Public Health, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins Bloomberg School of Public Health

14:50 A Unified Framework for Extracting Single-Cell Information from High-Resolution Spatial Transcriptomics Bayesian Analysis of Frechet Regression
Can Yang
Hong Kong University of Science and Technology

15:15 CoxFormer Enables Spatial Omics Inference with Multimodal Generative Modeling
⋆ *Jin Liu, Yiyang Yang*
The Chinese University of Hong Kong, Shenzhen, The Chinese University of Hong Kong, Shenzhen

S055: Innovations in Regression and Change Detection for Complex Data

Room: Chi Wah Lecture Hall 108
Organizer: Xinyuan Song

Chair: Xiangnan Feng

14:00 Cross-Semantic Transfer Learning for High-Dimensional Linear Regression
Xuejun Jiang
Southern University of Science and Technology

14:25 Censored Quantile Regression With Latent Variables
Haijin He
Shenzhen University

14:50 Bayesian Analysis of Frechet Regression
⋆*Hefei Liu, Yingchun Huang, Pengfei Liu, Xinyuan Song*
Yunnan University of Finance and Economics, Yunnan University of Finance and Economics, Jiangsu Normal University, The Chinese University of Hong Kong

15:15 Detection of Temporal Changes in Self-Exciting Point Processes
Jiayu Huang
Central China Normal University

S073: Modern Statistical Methods for Complex Data: Hypothesis Testing, Estimation and Application

Room: Chi Wah Lecture Hall 301
Organizer: Bingshu Chen
Chair: Bingshu Chen

14:00 Adaptive-Gated Testing Framework for Transfer Learning: Model Structural Integrity and Functional Similarity
⋆*Yu Peng, Jiaqi Huang, Falong Tan, Lixing Zhu*
Beijing Normal University at Zhuhai, University of Michigan, Hunan University, Beijing Normal University at Zhuhai

14:25 Leveraging Big Data: Practical Strategies in Large-Scale Prospective Cohort Studies
Simin Li, Zhang Chen, Wanru Yang, Mingyun Chen, ⋆ *Xinzhi Li*
Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology

14:50 E-Variable-Based Framework for High-Dimensional Change Point Detection
⋆*Feilong Hu, Jiaqi Huang, Guanxun Li, Lixing Zhu*
Beijing normal university at Zhuhai, University of Michigan, United States, Beijing normal university at Zhuhai, Beijing normal university at Zhuhai

15:15 Test for a Biomarker Threshold Effect in Generalized Linear Models with Application to Clinical Trials
⋆*Bingshu Chen, Wenyu Jiang*
Queen's University, Queen's University

S088: Precision Medicine and High-Dimensional Data Inference

Room: Chi Wah Lecture Hall 204
Organizer: Xuerong Chen
Chair: Xuerong Chen

14:00 Optimal Treatment Regimens in Clinical Studies of Colorectal Cancer
Zicheng Lin, ⋆ *Cunjie Lin*
Renmin University of China, Renmin University of China

14:25 Two-Way Subgroup Structure Learning for High Dimensional Clustered Survival Data
Ye He, Dongsheng Tu, Liu Liu, ⋆ *Ling Zhou*
Sichuan Normal University, Queen's University, Kingston, Canada, Sichuan Normal University, Southwestern University of Finance and Economics

14:50 A General Distributed Privacy-Preserving Statistical Inference Framework for Massive Data
Senlin Yuan, ⋆ *Xuerong Chen, Yong Zhou*
Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Southwestern University of Finance and Economics, East China Normal University

15:15 Hypothesis Testing in High-Dimensional Censored-Transformation Models
Xu Liu
Shanghai University of Finance and Economics

S097: Recent Advances in Data-Driven Decision Making

Room: Chi Wah Lecture Hall 205
Organizer: Lu Wang
Chair: Lu Wang

14:00 Data-Driven Price-Setting Newsvendor Problem
Ye Lu
City University of Hong Kong

14:25 Federated Learning with Heterogeneous Data: Challenges, Algorithms, and Practical Applications
Lei Shi
Hefei University of Technology

14:50 Adaptive Trial Design: Enabling Dynamic Decisions for Innovative Therapies
Ke Zhang
SCG Cell Therapy Pte. Ltd.

15:15 AI-Driven Personalization: Causal Learning + RL for Dynamic Treatment Decisions
Lu Wang
University of Michigan

S104: Recent Advances in Omics Data Analysis Techniques

Room: Chi Wah Lecture Hall 303
Organizer: Tianwei Yu
Chair: Tianwei Yu

14:00 Robust Statistical Methods for Analyzing Microbial Co-Abundance and Co-Occurrence Network
Yicong Mao, ⬩ *Yijuan Hu*
Peking University, Peking University

14:25 Computational Modeling of Spatial Transcriptomics: From Feature Representation and Cross-Modal Prediction to Statistical Inference
Hao Wu
Shenzhen University of Advanced Technology

14:50 A Bayesian Clustering Method for Spatial Omics Data Based on Factor Analysis
Yongyi Luo, Bencong Zhu, ⬩ *Xiaodan Fan*
The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

15:15 Keep Everything on the Same Page — A Novel Strategy for Omics Data Integrationintegration
Lixin Cheng
Shenzhen People's Hospital, the First Affiliated Hospital of Southern University of Science and Technology

S109: Recent Advances in Statistical Designs for Early-Phase Oncology Clinical Trials

Room: Chi Wah Lecture Hall 304
Organizer: Yong Zang
Chair: Yan Han

14:00 A Bayesian Pharmacokinetics-Guided Regimen Optimization with Delayed Toxicity and Survival Outcomes
Mengyi Lu, Guanfu Liu, ⬩ *Rongji Mu*
Nanjing Medical University, ShangShanghai University of International Business and Economics, Shanghai Jiao Tong University

14:25 Great Wall: A Generalized Dose Optimization Design for Drug Combination Trials Maximizing Survival Benefit
Yan Han
Indiana University

14:50 Precision Generalized Phase I-II Designs
Yong Zang
Indiana University School of Medicine

S114: Recent Advances on the Analysis of Complex Data

Room: Chi Wah Lecture Hall 210
Organizer: Jianguo Sun
Chair: Jianguo Sun

14:00 Novel Empirical Likelihood Method for the Cumulative Hazard Ratio Under Stratified Cox Models
Dazhi Zhao, ⬩ *Yichuan Zhao*
Georgia State University, Georgia State University

14:25 Variable Selection with Broken Adaptive Ridge Regression for Interval-Censored Competing Risks Data
Fatemeh Mahmoudi, Chenxi Li, Kaida Cai, ⬩ *Xuewen Lu*
Mount Royal University, Michigan State University, Southeast University, University of Calgary

14:50 Estimation of Stratified Semiparametric Transformation Models Under Interval Censoring with Auxiliary Information
⬩ *Yinghao Pan, Yanqing Sun, Xi Ning, Peter Gilbert*
University of North Carolina at Charlotte, University of North Carolina at Charlotte, Colby College, Fred Hutchinson Cancer Center

15:15 A Factor-Augmented Regression Approach to Learning Bivariate Interval-Censored Data
Shuwei Li
Guangzhou University

S126: Statistical and Computational Methods for Omics and Biological Discovery

Room: Chi Wah Lecture Hall 305
Organizer: Liangliang Zhang
Chair: Liangliang Zhang

14:00 Structured and Scalable AI for High-Dimensional Biological Data: Beyond Lasso Under Correlation
Xiaoqing Huang, Aolong Li, ⬩ *Yijie Wang*
Indiana University Bloomington, Indiana University Bloomington

14:25 Model-Based Multifacet Clustering with High-Dimensional Omics Applications
⬩ *Wei Zong, Danyang Li, Marianne Seney, Colleen Mcclung, George Tseng*
The Chinese University of Hong Kong, Shenzhen, University of Pittsburgh, University of Pittsburgh, University of Pittsburgh, University of Pittsburgh

14:50 Covariate-Informed Multiple Testing via the Conditional Local False Discovery Rate, with an Application to a Melanoma Study
Ziyi Wang, Satabdi Saha, Christine Peterson, ⬩ *Yushu Shi*
University of Texas Health, University of Texas MD Ander-

son Cancer Center, Rice University, Weill Cornell Medicine

S128: Statistical Foundations for Machine Learning and AI

Room: Chi Wah Lecture Hall 206
Organizer: Weijie Su
Chair: Xiang Li

14:00 Labels or Preferences? Budget-Constrained Learning with Human Judgments over AI-Generated Outputs
Zihan Dong, Xiaotian Hou, ⋆ *Ruijia Wu, Linjun Zhang*
Rutgers University, University of Pennsylvania, Shanghai Jiao Tong University, Rutgers University

14:25 Data Selection for LLM: Evolving from Closed to Open
Jun Shu
Xi'an Jiaotong University

14:50 Spatially Randomized Designs Can Enhance Policy Evaluation
Ying Yang
Fudan University

15:15 What Separates Useful from Useless Synthetic Data? Verifying Synthetic Data Through Representations
Chendi Wang
Xiamen University

S050: Frontiers of Transfer Learning, Data Fusion, and Adaptive Evidence Generation: Theory, Methods, and Practice

Room: Chi Wah Lecture Hall 207
Organizer: Kun Chen
Chair: Jianmin Chen

14:00 Targeted Learning via Probabilistic Subpopulation Matching
⋆ *Xiaokang Liu, Jie Hu, Naimin Jing, Yang Ning, Cheng Yong Tang, Runze Li, Yong Chen*
University of Missouri, University of Pennsylvania, Merck & Co., Inc, Cornell University, Temple University, Penn State University, University of Pennsylvania

14:25 Efficient Semiparametric Inference for Distributed Data with Blockwise Missingness
Jingyue Huang, ⋆ *Huiyuan Wang, Yuqing Lei, Yong Chen*
University of Pennsylvania, University of Pennsylvania, University of Pennsylvania, University of Pennsylvania

14:50 Surrogate-Powered Inference
⋆ *Jianmin Chen, Huiyuan Wang, Thomas Lumley, Xiaowu Dai, Yong Chen*
Xiamen University,University of Pennsylvania, University of Auckland, University of California, Los Angeles, University of Pennsylvania

15:15 A Unified Information–Time Paradigm for Single-Arm Time-to-Event Trials: The Window-Cohort with Calibrated Follow-Up Requirement
Haitao Pan
St. Jude Children's Research Hospital

S136: Statistical Learning and Modeling for Complex Data

Room: Chi Wah Lecture Hall 306
Organizer: Zheyang Wu
Chair: Donghui Yan

14:00 Utilizing Large Models in Biomedical Data Analysis: Some Case Studies
Yinglei Lai
George Washington University

14:25 Fuzzy Join of Data for Statistical Learning
⋆ *Donghui Yan, Yingjie Wang, Haiping Xu, Chenyu Li, Ryan Berry, Hongyu Zhao*
UMass Dartmouth, UMass Dartmouth, UMass Dartmouth, Columbia University, UMass Dartmouth, Yale University

14:50 Structural Learning of Arbitrary Graphs under Potts Model Framework
Xinran Bian
Sun Yat-Sen University

15:15 Covariate-Adjusted Win Statistics in Randomized Clinical Trials with Ordinal Outcomes
Zhiqiang Cao
Shenzhen Technology University

S147: Study Design and Joint Modeling with Longitudinal and Survival Data

Room: Chi Wah Lecture Hall 307
Organizer: Zhigang Li
Chair: Yang Yang

14:00 Longitudinal Variable Selection Under Joint Modeling of Longitudinal and Binary Data
⋆ *Fangfang Wang, Lihui Zhao*
Tianjin University of Finance and Economics, Northwestern University

14:25 Parsimonious Joint Model of Survival Outcome and Multiple Longitudinal Risk Factors
⋆ *Zhiyang Zhou, Lihui Zhao*
University of Florida

14:50 Integrating Intensive Mobile Monitoring Data for Dynamic Prediction of Healthcare Utilization via Landmark Models and FPCA
Qing Yang, Anushka Palipana, Wenbin Lu, Ryan Shaw, Matt Crowly
Duke University, Duke University, North Carolina State University, Duke University, Duke University

15:15 Regression-Based Modeling of Spearman's Rho for Longitudinal Metabolomics and Mental Wellness in Breast Cancer Patients
Tuo Lin
University of Florida

S154: When Large Language Model Meets Statistics

Room: Chi Wah Lecture Hall 208
Organizer: Jingyuan Liu
Chair: Jingyuan Liu

14:00 Evaluating Uncertainty of Large Language Models via Distributional and Structural Estimation
Pengkun Yang
Tsinghua University

14:25 When Kernel Ridge Regression Meets Large Language Models
Xin Bing, Xin He, Chao Wang
University of Toronto, Shanghai University of Finance and Economics, University of Toronto

14:50 Efficient Reasoning with Statistical Guarantees for Large Language Models
Hao Zeng, Jianguo Huang, Bingyi Jing, Hongxin Wei, Bo An
Southern University of Science and Technology

15:15 AlphaPareto: Formulaic Alpha Discovery with LLM-Guided Multi-Objective Reinforcement Learning
Yingbo Zhao, Zeyu Yang, Zhoufan Zhu
Xiamen University, Xiamen University, Xiamen University

S005: Advanced Statistical Methods for Complex Survival Data

Room: Chi Wah Lecture Hall 308
Organizer: Peijie Wang
Chair: Peijie Wang

14:00 Distributed Censored Quantile Regression: Convolution Smoothing and Communication Efficiency
Ying Huang, Xiaoqi Jiao, Huijuan Ma, Baoxue Zhang
East China Normal University

14:25 Regression Analysis of Dependent Current Status Data with Left-Truncation Under Linear Transformation Model
Mengyue Zhang, Shishun Zhao, Da Xu, Tao Hu, Jianguo Sun
Jilin University, Jilin University, Northeast Normal University, Capital Normal University, Southern University of Science and Technology

14:50 Dynamic Survival Prediction Based on Multiple Intermediate Events
Tonghui Yu, Liming Xiang
Harbin Institute of Technology, China, Nanyang Technological University, Singapore

15:15 Online Updating and Heterogeneity Detection for Additive Hazard Models with Streaming Survival Data
Lina Wang, Zhanhang Gao, Bing Wang
Anhui University, Anhui University, Anhui University

b. June 27th PM (16:00-17:40)

S021: Advances in Statistical Modeling for Causal Inference and High-dimensional Complex Data

Room: Chi Wah Lecture Hall 304
Organizer: Peijie Wang
Chair: Peijie Wang

16:00 Estimating Conditional Complier Quantile Treatment Effect Via Stratified Quantile Regression
Huijuan Ma, Mengjiao Peng, Jing Qin
East China Normal University, East China Normal University, National Institute of Allergy and Infectious Diseases, National Institutes of Health

16:25 Optimal Dynamic Treatment Regimes for High-Dimensional Accelerated Failure Time Model
Shijie Zhao, Wei Zhao
Shandong University, Shandong University

16:50 Variable Selection for Mixed Panel Count Data Using Gaussian Seamless-L0 Penalty
Lei Ge, Tao Hu, Shurong Zheng, Jianguo Sun
Northeast Normal University, Capital Normal University, Northeast Normal University, Southern University of Science and Technology

17:15 Regression Analysis of Multiplicative Hazards Model with Time-Dependent Coefficient for Sparse Longitudinal Covariates
Zhuowei Sun, Hongyuan Cao
Dalian Medical University, Florida State University

S025: AI and Big Data Analysis

Room: Chi Wah Lecture Hall 102
Organizer: Yichuan Zhao
Chair: Ting Zhang

16:00 Global Inference for Distribution Functions in Large-Scale Data: A Bootstrap Approach with Distributed Computation
Li Cai
Zhejiang Gongshang University

16:25 Adaptive Allocation Method for Large Portfolio
Haitao Tian, Shouri Hu, ⋆ *Lei Huang*
Southwest Jiaotong University

16:50 Copula-Enhanced Vision Transformer for High Myopia Diagnosis Through OU UWF Fundus Images
Chong Zhong, Yunhao Liu, Yang Li, Xiang Fu, Jin Yang, Danjuan Yang, Meiyan Li, Xingtao Zhou, Bo Fu, ⋆ *Catherine Liu*
The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, Fudan University, Fudan University, The Hong Kong Polytechnic University, Eye & ENT Hospital, Fudan University, Eye & ENT Hospital, Fudan University, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

17:15 PB-DiffHiC: a Statistically Principled Method for Detecting Differential Chromatin Interactions Using raw Pseudo-Bulk Hi-C Data
⋆ *Yan Zhou, Yaohua Hu, Liuting Tan, Jiadi Zhu, Yutong Fei, Ming Gu, Dechao Tian*
Shenzhen University, Shenzhen University, Shenzhen University. Shenzhen University, Shenzhen University. Sun Yat-sen University (Shenzhen Campus),Sun Yat-sen University (Shenzhen Campus)

S030: Bootstrap and Model Averaging Methods for Robust Statistical Learning

Room: Chi Wah Lecture Hall 208
Organizer: Chenglong Ye
Chair: Xiaofang Shi

16:00 Bootstrap Model Averaging
⋆ *Minghui Song, Guohua Zou, Alan T.K. Wan*
Beijing Wuzi University, Capital Normal University, City University of Hong Kong

16:25 Model Screening in Model Averaging: How and When?
⋆ *Guangyuan Cui, Alan Wan, Xinyu Zhang*
City University of Hong Kong Shenzhen Research Institute, City University of Hong Kong, Academy of Mathematics and Systems Science

16:50 High-Resolution Feature Identification in High-Dimensional Clustering
Lyuou Zhang
Shanghai University of Finance and Economics

17:15 Sign Hacking in The Age of Big Data
⋆ *Jie Hu, Yuhong Yang*
Tsinghua University, Tsinghua University

S041: Advances in Statistical Inference for Dependent and High-Dimensional Data

Room: Chi Wah Lecture Hall 103
Organizer: Xuejun Jiang
Chair: Jiancheng Jiang

16:00 Precise Asymptotics and Refined Regret of Variance-Aware UCB
⋆ *Yingying Fan, Yuxuan Han, Jinchi Lv, Xiaocong Xu, Zhengyuan Zhou*
University of Southern California

16:25 Fast Community Detection of Discrete-Time Temporal Networks in Mixture Dynamic Stochastic Block Models
Yang Liu, ⋆ *Binghui Liu*
Northeast Normal University, Northeast Normal University

16:50 Statistical Inference for Power Autoregressive Conditional Duration Models with Stable Innovations
⋆ *Yuxin Tao, Dong Li*
Southern University of Science and Technology, Tsinghua University

S045: Frontiers in Biostatistics: Robust Inference and Machine Learning

Room: Chi Wah Lecture Hall 104
Organizer: Jialiang Li
Chair: Mu Yue

16:00 Heterogeneous Latent Variable Modeling of Multivariate Outcomes
Yi Liu, John Hanfelt, ⋆ *Limin Peng*
Emory University, Emory University, Emory University

16:25 Bilevel Gene Network Estimation with a Deep Neural Network Under Hierarchical Structured Sparsity
Jiayi Fan, Jingyuan Yang, Shuangge Ma, ⋆ *Mengyun Wu*
Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Yale School of Public Health, Shanghai University of Finance and Economics

16:50 On Multiple Robustness of Proximal Dynamic Treatment Regimes
⋆ *Yuanshan Gao, Yang Bai, Yifan Cui*
Zhejiang University, National University of Singapore, Zhejiang University

17:15 A Machine Learning-Based Variance Reduction Approach of Average Treatment Effect Estimation in Observational Studies
Hao Mei, Siqi Liu, ⬩ *Zhibo Cai*
Renmin University of China, Renmin University of China, Renmin University of China

S129: Statistical Foundations for Preference Learning, Representation, and Heterogeneity in the Era of AI

Room: Chi Wah Lecture Hall 106
Organizer: Yong Chen
Chair: Jianmin Chen

16:00 When Less Is More: Binary Feedback Can Outperform Ordinal Comparisons in Ranking Recovery
⬩ *Shirong Xu, Jingnan Zhang, Junhui Wang*
Xiamen University, University of Science and Technology of China, The Chinese University of Hong Kong

16:25 Stein-Encoder: A White-Box Supervised Encoder via Stein Identities in Multi-Modal Studies
Jiarui Zhang, Shuoxun Xu, ⬩ *Jiasheng Shi, Xinzhou Guo*
The Chinese University of Hong Kong, Shenzhen, The Hong Kong University of Science and Technology

16:50 Learning Coefficient Heterogeneity with Hierarchical Structure Information
⬩ *Yan Zhong, Qiao Zheng, Lei Ma, Jianhua Huang*
East China Normal University, Shanghai University of International Business and Economics, East China Normal University, The Chinese University of Hong Kong, Shenzhen

17:15 A Statistical Framework for Alignment with Biased AI Feedback
⬩ *Xintao Xia, Zhiqiu Xia, Linjun Zhang Zhang, Zhanrui Cai*
Zhejiang University, Rutgers University, The University of Hong Kong

S054: Inference and Learning Under Heterogeneous Data

Room: Chi Wah Lecture Hall 305
Organizer: Nan Lin
Chair: Nan Lin

16:00 Online Inference Under Over-Parameterized Models with Hidden Confounders
Shuyan Chen, ⬩ *Xingdong Feng, Yeheng Ge, Tao Li, Mengyun Wu*
University of Science and Technology, Shanghai University of Finance and Economics, The Hong Kong Polytechnic University, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

16:25 Heterogeneous Quantile Treatment Effect Estimation for Longitudinal Data with High-Dimensional Confounding
Zhixin Qiu, Huichen Zhu, Wenjie Wang, ⬩ *Yanlin Tang*
East China Normal University, The Chinese University of Hong Kong, Eli Lilly and Company, East China Normal University

16:50 A Transfer Learning Framework for Efficient Portfolio Selection
Hongmei Lin
Shanghai University of International Business and Economics

17:15 Federated R-Learner for Heterogeneous Treatment Effect Estimation
Runyan Xin, Yiqiao Jin, ⬩ *Nan Lin, Linying Zhang*
Washington University in St. Louis, Washington University in St. Louis, Washington University in St. Louis, Washington University in St. Louis

S056: Innovative Approaches in Machine Learning and Health Data Analytic

Room: Chi Wah Lecture Hall 107
Organizer: Chen-An Tsai
Chair: Chung Chang

16:00 Cognition-Guided Regional Amyloid Thresholds Detect Earlier Domain-Specific Decline in Cognitively Normal Adults
Chung Chang
Sun Yat-sen University, Taiwan

16:25 An Interpretable Hybrid Framework: Integrating Probabilistic Generative Modeling with Weighted Logistic Regression for Imbalanced Biomedical Data
⬩ *Charlotte Wang, Yu-Ting Huang*
Taiwan University, Taiwan University

16:50 A Prediction Platform to Estimate the Phytohormone Synthesis Potential in Bacteria Using Genomic Data
Zheng-Xiang Ye, ⬩ *Steven Hung-Hsi Wu*
Taiwan University, Taiwan University

17:15 A Data-Informed Variational Clustering Framework for Noisy High-Dimensional Data
Wanping Chen
Fu Jen Catholic University

S058: Innovative Statistical Design Methods for Modern Clinical Trials

Room: Chi Wah Lecture Hall 108
Organizer: Shouhao Zhou
Chair: Yuexi Wang

16:00 Posterior Predictive (PoP) Design for Next-Generation Phase I Clinical Trials
Chenqi Fu, ⬩Shouhao Zhou, J Jack Lee
Pennsylvania State University, Pennsylvania State Univer -sity, UT MD Anderson Cancer Center

16:25 A Pharmacokinetics-Informed Bayesian Design for Dose Optimization in Multi-Regional Clinical Trials
Xin Chen,Wenyun Yang,⬩Fangrong Yan
China Pharmaceutical University, China Pharmaceutical University, China Pharmaceutical University

16:50 Metaheuristics for Finding Optimal Designs for Dose Response Trials
Weng-Kee Wong
University of California at Los Angeles

17:15 Learning from Literature: Integrating LLMs and Bayesian Hierarchical Modeling for Next-Generation Clinical Trial Design
Wei Wei
Yale University

S059: Integrative Statistical Methods for Genetic Association and Multi-Omics Studies

Room: Chi Wah Lecture Hall 202
Organizer: Xiang Zhou
Chair: Xiang Zhou

16:00 Modeling and Predicting Single-Cell Multi-Gene Perturba-tion Responses
Gefei Wang
Yale University

16:25 Novel Multi-Context Transcriptome-Wide Association Framework Using Penalized Tensor Regression
Ruilong Chen, Qiang Liu, Michael Epstein, ⬩Jingjing Yang
Emory University, Emory University, Emory University, Emory University

16:50 Multi-resolution Clustering with Cross-Algorithm Reconciliation in Genomic Data Analysis
⬩Xiaoquan Wen, Ali Turfah
University of Michigan, University of Michigan

S075: Modern Statistical Methods for High-Dimensional and Structured Data

Room: Chi Wah Lecture Hall 105
Organizer: Xinyuan Song
Chair: Kai Kang

16:00 Testing for Change Intervals in High-Dimensional Time Series
⬩Kin Wai Chan, Ting Tin Ma
The Chinese University of Hong Kong, The Chinese University of Hong Kong

16:25 Spatial Varying Coefficient Model with Covariate Measurement Errors: Bias-Corrected Inference and Efficient Computation
Bohai Zhang
Beijing Normal-Hong Kong Baptist University

16:50 Towards Understanding Massive Activations in Attention Sink Mechanism
Yuanyuan Lin
The Chinese University of Hong Kong

17:15 Errors-In-Variables Gaussian Processes for Mixed-Input Regression
Penghui Fu, ⬩ Sheng Jiang
The Chinese University of Hong Kong, Shenzhen, The Chinese University of Hong Kong, Shenzhen

S076: Modern Statistical Methods for High-Dimensional Data and Machine Learning Models

Room: Chi Wah Lecture Hall 112
Organizer: Quefeng Li
Chair: Zhao Chen

16:00 Policy Evaluation in Distributional Reinforcement Learning via Kernel Mean Embeddings
Mehrdad Mohammadi, Qi Zheng, ⬩Ruoqing Zhu
University of Illinois Urbana Champaign, University of Louisville, University of Illinois Urbana Champaign

16:25 On Reference-Regulated Multiperiod Mean-Variance Portfolio Optimization in High Dimensions
Yutao Deng, Jianjun Gao, ⬩Weichen Wang
The University of Hong Kong

17:50 Multi-View Orthogonal Projection Regression for Multi-Omics Data
Zongrui Dai, Yvonne Huang, ⬩Gen Li
University of Michigan, University of Michigan, University of Michigan

S079: New Advances in Infectious Disease Modeling

Room: Chi Wah Lecture Hall 203
Organizer: Zhigang Li
Chair: Zhigang Li

16:00 Improving Assessment of Vaccine Effectiveness by Coupling Test-Negative Design Studies with Survival Models
Shangchen Song, Matt Hitchings, ⬩Yang Yang, Ira Longini
University of Georgia

16:25 Estimating Real-Word Population-Level Vaccination Effects
Ryan Jasper, Nicholas Seewald, ⋆ *Jing Huang*
University of Pennsylvania, University of Pennsylvania, University of Pennsylvania

16:50 Statistical and Machine Learning Approaches for Infectious Disease Dynamics
Tim Tsang
University of Hong Kong

17:15 A Multiscale Spatio-Temporal Data Monitoring Framework for Prospective Infectious Disease Surveillance
⋆ *Kai Yang, Peihua Qiu*
Medical College of Wisconsin, University of Florida

S107: Recent Advances in Spatial Omics and Single-cell Data Analysis

Room: Chi Wah Lecture Hall 306
Organizer: Yingying Wei
Chair: Fangda Song

16:00 Graph Transformer Inference Links Ligand-Receptor Signaling to Target Genes in Special Niches
Xiao Xiao, Yihuan Wang, Hongyu Zhao, Grace Bushnell, ⋆ *Zuoheng Wang*
Yale University, Yale University, Yale University, University of Minnesota, Yale University

16:25 Background-Correction of Spatial Proteomics Data Using Finite Mixture of Gaussian Models
⋆ *Agus Salim, Dharmesh Bhuva, Malvika Kharbanda*
University of Melbourne, University of Queensland, Adelaide University

16:50 Multi-Scale Learning Spatially Interpretable Latent Feature Representations in Spatial Transcriptomics
⋆ *Yuanhua Huang, Jun Wang, Ole Winther, Tianjie Wang, Lingyu Li*
University of Hong Kong, University of Copenhagen, University of Copenhagen, University of Hong Kong, University of Hong Kong

17:15 BayesRare: A Bayesian Mixture Model for Population-Level Rare Cell Type Detection in Multi-Subject Single-Cell RNA Sequencing Data
⋆ *Yinqiao Yan, Hao Wu*
Beijing University of Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences; Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology

S112: Recent Advances of Time Series Analysis and Change-Point Detection

Room: Chi Wah Lecture Hall 201
Organizer: Zhanrui Cai
Chair: Yuxin Tao

16:00 Robust and Efficient Inference for Nonstationary GARCH(1,1) Models with Heavy-Tailed and Asymmetric Innovations
⋆ *Huan Gong, Dong Li, Yuxin Tao*
National University of Defense Technology, Tsinghua University, Southern University of Science and Technology

16:25 Adaptive Dynamic Pricing Under Non-Stationary Environments
⋆ *Feiyu Jiang, Zifeng Zhao*
Fudan University, University of Notre Dame

16:50 Two-Way Matrix Autoregressive Model with Thresholds
Cheng Yu, ⋆ *Xinyu Zhang, Dong Li, Howell Tong*
The University of Chicago, East China Normal University, Tsinghua University, London School of Economics and Political Science

17:15 Ranking Individuals in Team Tournaments: With an Application to E-Sport Games
⋆ *Xuanling Yang, Yuxin Tao, Zixuan Sheng, Dong Li*
Beijing University of Chemical Technology, Southern University of Science and Technology, Tsinghua University, Tsinghua University

S117: Recent Developments on Applied Statistics

Room: Chi Wah Lecture Hall 301
Organizer: Tsung-I Lin
Chair: Chang-Yun Lin

16:00 Model-Robust Process Optimization for Data with a Multi-Stratum Structure
Xiaohua Liu, ⋆ *Po Yang, Chang-Yun Lin*
University of Manitoba, University of Manitoba, Chung Hsing University

16:25 Shannon's Entropy Estimation Under Undersampling
Tsung-Jen Shen
Chung Hsing University

16:50 Clustering High-Dimensional Data with Missing Values Using Restricted Multivariate Skew-t Mixtures of Common Factor Analyzers
⋆ *Tsung-I Lin, Wan-Lun Wang, Luis M. Castro, Shi-Xiu Yu*
Chung Hsing University, Cheng Kung University, Pontificia Universidad Católica de Chile, Chung Hsing University

17:15 Efficient and Robust Block Designs for Order-of-Addition Experiments
Chang-Yun Lin
Chung Hsing University

S122: Some New Methods of Robust Statistical Learning for Complex Data

Room: Chi Wah Lecture Hall 204
Organizer: Hui Zhao
Chair: Xiaofei Zhang

16:00 Local linear QR estimator of conditional quantile for strong-mixing and truncated data
Mei Yao,Jiangfeng Wang,Lu Lin, Yu Xia
Hefei University of Technology,Zhejiang Gongshang University,Shandong University, Hangzhou Normal University

16:25 Robust Subgroup Analysis for Heterogeneous Censored Data
Zhaohui Xu, Daoji Li, Zemin Zheng
University of Science and Technology of China, California State University, Fullerton, University of Science and Technology of China

16:50 Stable Semi-Distance Covariance for Independence Testing: From Low to High Dimensions
Feng Zou
Zhongnan University of Economics and Law

17:15 Robust Reproducibility Method Based on Modal Regression
Jinzhao Yu, Hui Zhao, Daoji Li
Hubei University of Economics

S124: Statistical Analysis of Complex Data

Room: Chi Wah Lecture Hall 205
Organizer: Xinyu Zhang
Chair: Ziwen Gao

16:00 Combining Pre-Trained Models Via Localized Model Averaging
Ziwen Gao
Tsinghua University

16:25 Frechet Regression with Mondrian Forests: Finite-Sample Guarantees and Ensemble Benefits
Rui Qiu, Fang Yao, Zhou Yu
Peking University, Peking University, East China Normal University

16:50 Bias-Corrected Weighted Least Squares Model Averaging in Linear Measurement Error Models With Right Censored Data
Guozhi Hu
Hefei Normal University

17:15 Supervised Functional Tensor Autoregression
Songhua Tan
University of Science and Technology of China

S127: Statistical and Computational Methods for Single-Cell, Spatial, and Multi-Omics Integration

Room: Chi Wah Lecture Hall 210
Organizer: Hongkai Ji
Chair: Hongkai Ji

16:00 Designing Spatial Transcriptomics Using Histology-Driven Expression Prediction
Jean Yee Hwa Yang
The University of Sydney

16:25 IntegrateRigor: Statistical Framework for Rigorous Integration of Single-Cell and Spatial Transcriptomic Data Without Prior Annotations
Jingyi Jessica Li
Fred Hutchinson Cancer Center

16:50 FastCCC: A Permutation-Free Framework for Scalable, Robust, and Reference-Based Cell-Cell Communication Analysis in Single Cell Transcriptomics Studies
Siyu Hou, Wenjing Ma, Xiang Zhou
Yale University, Yale University, Yale University

17:15 Computational Methods and Resources for Multi-Omics Understanding of Complex Human Systems
Yun Li, Jiawen Chen, Didong Li
University of North Carolina, Chapel Hill, University of North Carolina, Chapel Hill, University of North Carolina, Chapel Hill

S130: Statistical Foundations of LLM Behavior, Alignment, and Automated Reasoning

Room: Chi Wah Lecture Hall 206
Organizer: Chenglong Ye
Chair: Chenglong Ye

16:00 Divergence Representations of LLM Safety Alignment
Qifan Song, Yue Xing, Rajdeep Haldar, Ziyi Wang, Guang Lin
Purdue University, Michigan State University, Purdue University, Purdue University, Purdue University

16:25 Transformers Are Born Biased: Structural Inductive Biases and Their Practical Consequences
Li Siquan, Tong Yao, Wang Haonan, Tianyang Hu
The Chinese University of Hong Kong, Shenzhen, National University of Singapore, National University of Singapore, The Chinese University of Hong Kong, Shenzhen

16:50 Theoretical Modeling of Large Language Model Self-Improvement Training Dynamics Through Solver-Verifier Gap
Yifan Sun, Yushan Liang, Zhen Zhang, Xin Liu, Jiaye Teng
Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Shanghai University

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

17:15 Qualified LLM Judge Aggregation with Calibrated Uncertainty
Yunan Wu
Tsinghua University

S131: Statistical Frontiers in Spatial Omics

Room: Chi Wah Lecture Hall 207
Organizer: Jingyi Jessica Li
Chair: Guan' ao Yan

16:00 Intelligent Integration Methods for Single-Cell Chromatin Accessibility Sequencing Data
Heyang Hua, Shengquan Chen
Nankai University, Nankai University

16:25 Benchmarking Cell-Type-Specific Spatially Variable Gene Detection Methods Using a Realistic and Decomposable Simulation Framework
Weiqi Li, Xinzhou Ge, Yuan Jiang
Oregon State University, Oregon State University, Oregon State University

16:50 Conditional Semi-Distance Correlation: Measuring Conditional Dependence in Spatial Transcriptomics
Ziqing Guo, Yang Feng, Jingyi Jessica Li, Lucy Xia
Hong Kong University of Science and Technology, New York University, Fred Hutchinson Cancer Center, Hong Kong University of Science and Technology

17:15 MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning
Yuying Xie, Jingru Guo, Yan Cui, Hongzhi Wen, Rob Yang
Michigan State University, Michigan State University, CartaBio Inc, Cartabio Inc, Cartabio Inc

S140: Statistical Methods and Applications

Room: Chi Wah Lecture Hall 302
Organizer: Jianguo Sun
Chair: Jianguo Sun

16:00 On Novel Applications of Classical Linear Regression in Biomedical Research
Zhezhen Jin
Columbia University

16:25 Meta- vs. Mega-Analysis: A Long-Standing Debate in Omics Data Integration
Jian Zou
Chongqing Medical University

16:50 Preserving Rare Features in Big Data Regression: Balanced Subsampling
Qingkai Dong, Jun Yan, Qiang Zhang, HaiYing Wang
University of Connecticut, University of Connecticut, Wills Eye Hospital, University of Connecticut

17:15 Simplifying Inputs of Machine Learning-Based Risk Calculators to Improve Clinical Usability
Zhe Lu, Kenneth Liao, Zhiyu Kang, Linwan Xu, Cindy Im, Yan Yuan
University of Alberta, University of Minnesota, University of Minnesota, University of Alberta, University of Minnesota, University of Alberta

S148: Survival Analysis and Related Statistical Methods

Room: Chi Wah Lecture Hall 307
Organizer: Tao Hu
Chair: Yijun Wang

16:00 Simultaneous Variable Selection and Estimation for a Partially Linear Cox Model
Tingting Cai, Mengqi Xie, Tao Hu, Jianguo Sun
Capital Normal University, Capital Normal University, Capital Normal University, University of Missouri

16:25 Probit Time-to-Event Regression for Misclassified Group Testing Data
Lijun Fang, Tao Hu, Shuwei Li, Lianming Wang, Christopher S. McMahan, Joshua M. Tebbs
Capital Normal University, Capital Normal University, Guangzhou University, University of South Carolina, Clemson University, University of South Carolina

16:50 Distance Weighted Learning for Individualized Treatment Rules
Wenying Wang, Jichang Yu, Yuanshan Wu
Zhongnan University of Economics and Law

17:15 Fusion-Based Sparse Estimation for Current Status Data with Heterogeneous Auxiliary Information
Yijun Wang, Jie Li
Zhejiang Gongshang University

S106: Recent Advances in Sequential Methodologies

Room: Chi Wah Lecture Hall 308
Organizer: Yiming Xing
Chair: Ruizhi Zhang

16:00 Intelligence-Driven Quality Analytics: Integrating Generative AI with Statistical Process Control
Fugee Tsung
Hong Kong University of Science and Technology

16:25 Nearest-Neighbor Methods for Sequential Change-Point Detection in High-Dimensional and Non-Euclidean Data
Hao Chen
University of California, Davis

16:50 Sequential Multistream and Multihypothesis Testing With Asynchronous Decisions and Generalized Error Metrics
Yiming Xing
Tongji University

17:15 Optimal Bounded-Influence Procedure for Robust Sequential Change-Point Detection
Ruizhi Zhang
University of Georgia

S012: Advances in Machine Learning Methods for Survival and Event-Time Data

Room: Chi Wah Lecture Hall 303
Organizer: Gang Li
Chair: Wen Su

16:00 DH-AFT: Distributed, Heterogeneity-Aware AFT Models for High-Dimensional Multi-Center Survival Data
Hong Wang
Central South University

16:25 Improving Prediction for Interval-Censored Data Using Auxiliary Predictive Information
Jingyi Hu, ⋆ *Yu Gu*
University of Hong Kong, University of Hong Kong

16:50 Deep Conditional Generative Learning for Optimal Individualized Treatment Rules
⋆ *Xiangbin Hu, Wen Su, Zhisheng Ye, Xingqiu Zhao*
Beijing Institute of Technology, City University of Hong Kong, National University of Singapore, The Hong Kong Polytechnic University

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

S002: Advanced Methods for Complex Data Analysis

Room: Chi Wah Lecture Hall 302
Organizer: Po Yang
Chair: Po Yang

14:00 Revisiting the Alpha-Folding Multivariate Normal Distribution
Xiaoke Qin, Saman Muthukumarana, ⋆ *Kevin McGregor*
University of Manitoba, University of Manitoba, University of Manitoba

14:25 Bayesian Approaches for Multi-Class Fault Detection Classification
Saman Muthukumarana
University of Manitoba

14:50 Thompson Sampling with Bayesian Additive Regression Trees
Ruizhe Deng, Bibhas Chakraborty, Ran Chen, ⋆ *Yan Shuo Tan*
National University of Singapore, Duke-NUS Medical School, Washington University in St. Louis, National University of Singapore

15:15 Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach
Lei Ding
University of Manitoba

S092: Recent Advances for the Analysis of Complex Data

Room: Chi Wah Lecture Hall 102
Organizer: Mingyue Du
Chair: Mingyue Du

14:00 Functional Feature Screening with False Discovery Rate Control for Ultrahigh-Dimensional Longitudinal Data
Yexin Zhang
Wuhan University

14:25 Data-Driven Newsvendor Decision-Making: Newsvendor Representation Transfer Learning
Baihua He
University of Science and Technology of China

14:50 Regression Analysis of Recurrent Gap Times with Informative Observation and Censoring Times
⋆ *Shu-Hui Chang, Hsin-Yang Lu*
Taiwan University, Taiwan University

S014: Advances in Modern High-dimensional Statistics

Room: Chi Wah Lecture Hall 103
Organizer: Anru Zhang
Chair: Haoxuan Sun

14:00 A Frequency-Domain Approach for Integrating Multiple Functional Time Series
Zerui Guo, Jianbin Tan, ⋆ *Hui Huang*
Sun Yat-sen University, Duke University, Renmin University of China

14:25 Triple-Dyad Ratio Estimation for the p1 Model
Qunqiang Feng, Yaru Tian, ⋆ *Ting Yan*
University of Science and Technology of China, University

of Science and Technology of China, Central China Normal University

14:50 Covariate-Dependent Latent Gaussian Graphical Models for Mixed-Type Data
Yuyang Liu, Liuyi Lan, ⬩ Shan Luo, Yiwang Zhou, Chenlei Leng
Shanghai University of International Business and Economics, Shanghai Jiao Tong University, Shanghai Jiao Tong University, St. Jude Children's Research Hospital, Hong Kong Polytechnic University

15:15 Decoupling Regularization and Privacy in Differentially Private Ridge Regression and ERM
⬩ *Wanjie Wang, Tathagata Banjernee*
National University of Singapore, National University of Singapore

S022: Advances in Survival Analysis and Causal Inference

Room: Chi Wah Lecture Hall 105
Organizer: Xingqiu Zhao
Chair: Xingqiu Zhao

14:00 Inference and Variable Selection for Two-Phase Studies with High-Dimensional Covariates
Haoyang Wang, Qingning Zhou, ⬩ Kin Yau Wong
The Hong Kong Polytechnic University, The University of North Carolina at Charlotte, The Hong Kong Polytechnic University

14:25 Semiparametric Causal Inference for Right-Censored Outcomes with Many Weak Invalid Instruments
Qiushi Bu, ⬩ Wen Su, Zhonghua Liu, Xingqiu Zhao
City University of Hong Kong, City University of Hong Kong, Columbia University, The Hong Kong Polytechnic University

14:50 Deep Nonparametric Inference for Interval-Censored Data
◆ *Qiang Wu, Wen Su, Shuangge Ma and Xingqiu Zhao*
The Hong Kong Polytechnic University

S019: Advances in Statistical Methods for Clinical Trials: Longitudinal Data and Complex Data Mechanisms

Room: Chi Wah Lecture Hall 303
Organizer: Wei Tu
Chair: Wei Tu

14:00 Uncovering Hormone Trajectory Subtypes in the Menopausal Transition via Envelope–Regularized Multivariate Longitudinal Clustering
⬩ *Zihang Lu, Bei Jiang*
University of Toronto

14:25 Inverse Probability of Treatment Weighting for Estimation of Time Dependent Treatment Effects Under the Test-Negative Design
⬩ *Helen Bian, Cong Jiang, Denis DenisTalbot, Robert Platt, Mireille Schnitzer*
McGill University, Harvard University, Laval University, McGill University, University of Montreal

14:50 Disease Progression as a Common Driver of HRQoL Trajectories and Observation Schedules in Advanced Cancer Trials: A Progression-Stratified Pattern-Mixture Analysis
⬩ *Wensha Zhang, Wei Tu*
Queen's Univerisity, Queen's University

15:15 Loss to Follow-Up in Phase III Cancer Trials: Patient-Level Predictors and Impact on Treatment Effect Estimates
⬩ *Wei Tu, Nick Cruickshanks, Harriet Richardson*
Queen's University and Canadian Cancer Trials Group, Queen's University, Queen's University and Canadian Cancer Trials Group

S029: Big Data and AI

Room: Chi Wah Lecture Hall 104
Organizer: Mei-Ling Ting Lee
Chair: Chunling Liu

14:00 Flexible Semiparametric Modeling With Application to Causal Inference
Xingqiu Zhao
The Hong Kong Polytechnic University

14:25 A Deep Learning Prognostic Model Derived from Whole-Transcriptome Data in Lung Adenocarcinoma
Hsuan Yu Chen
Institute of Statistical Sciences, Academia Sinica, Taiwan

14:50 Domain-Specific Nonparametric Regression for Domain Generalization
⬩ *Siming Zheng, Yuanyuan Lin, Yong Zhou, Jian Huang*
Southeast University, The Chinese University of Hong Kong, East China Normal University, The Hong Kong Polytechnic University

S040: Experimental Design and Optimization

Room: Chi Wah Lecture Hall 304
Organizer: Fasheng Sun
Chair: Fasheng Sun

14:00 A General (Non-Markovian) Framework for Covariate Adaptive Randomization: Achieving Balance While Eliminating the Shift
Hengjia Fang, ⬩ Wei Ma

Institute of Statistics and Big Data, Renmin University of China

14:25 A New Discrepancy-Driven Framework for Uniform Factorial Designs with Application to Bayesian Structure Learning
Sumin Wang
Hebei University of Technology

14:50 Uniform Projection Designs Under the Stratified L2-Discrepancy
Sixu Liu, ⋆ *Yaping Wang*
Beijing Institute of Mathematical Sciences and Applications, East China Normal University

15:15 Optimal Design for Online Sequential Multi-Factorial Experiments with Optional Stopping
Xin Liu, ⋆ *Jun Yu, Wenbiao Zhao, Rui Zeng*
Donghua University, Beijing Institute of Technology, Beijing Institute of Technology, Beijing Institute of Technology

S052: High-Dimensional Inference and P-Value Combination

Room: Chi Wah Lecture Hall 112
Organizer: Qizhai Li
Chair: Haili Zhang

14:20 Jackknife Model Averaging with Quantile Regression Under Distributed Heterogeneous Data
Haili Zhang
Shenzhen Polytechnic University

14:25 Statistical Inference of the Random Coefficients Multinomial Logit Model
⋆ *Zixuan Wang, Jing Zhou, Junhao Hu, Daijun Wei, Qizhai Li*
South-Central Minzu University, Chinese Academy of Sciences, South-Central Minzu University, Hubei Minzu University, Chinese Academy of Sciences

14:50 Feature Screening for Multiple Responses
⋆ *Zhenzhen Jiang, Hongping Guo, Jinjuan Wang*
Hunan University of Technology and Business, Hubei Normal University, Beijing Institute of Technology

S062: Machine Learning and Statistical Modeling Applications

Room: Chi Wah Lecture Hall 106
Organizer: Fasheng Sun
Chair: Fasheng Sun

14:00 TSPINN: Thompson Sampling-Based Adaptive Training for Physics-Informed Neural Networks
Dianpeng Wang
Beijing Institute of Technology

14:25 Influence-Guided Subsampling Under Measurement Constraints
Lin Wang
Purdue University

14:50 Mesh-Clustered Gaussian Process Emulator for Partial Differential Equation Systems
Chih-Li Sung, ⋆ *Wenjia Wang, Liang Ding, Xingjian Wang*
Michigan State University, National University of Singapore, Fudan University, Tsinghua University

15:15 Noise Resistant Control Charts for Detecting Periodicity from Correlation
Yongxiang Li, Yunji Zhang, ⋆ *Qian Xiao, Jianguo Wu*
Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Peking University

S063: Microbiome Intelligence: Large Models, Big Data, and Statistical Advances

Room: Chi Wah Lecture Hall 202
Organizer: Liangliang Zhang
Chair: Ni Zhao

14:00 BGOB: A Deep-Learning-Based Interpolation of Longitudinal Microbiome Data
⋆ *Di Wu, Yixiang Qu, Ruiqi Lv*
University of North Carolina at Chapel Hill, University of North Carolina at Chapel Hill, Carnegie Mellon University

14:25 Human Gut Mycobiome and Health in Large-Scale Cohort Studies
Ju-Sheng Zheng
Westlake University

14:50 Integrative Statistical Analysis for Enhanced Scientific Discoveries in Metagenomics Studies
Xiang Zhan
Southeast University, School of Statistics and Data Science

15:15 Application of Artificial Intelligence Content Generation (AIGC) in Big Data Mining of Microbiome
⋆ *Kang Ning, Yuan Wang*
Huazhong University of Science and Technology, Huazhong University of Science and Technology

S069: Modern Statistical Inference and Modeling for Complex and Structured Data

Room: Chi Wah Lecture Hall 203
Organizer: Fengting Yi
Chair: Fengting Yi

14:00 A Study on the Performance of Matrix Control Charts Based on the Normal Distribution
Cong Li
Jilin University

14:25 Regression Analysis of Interval-Censored Failure Time Data with Change Points and a Cured Subgroup
Yichen Lou,◆ Mingyue Du, Xinyuan Song
The Chinese University of Hong Kong, Jilin University, The Chinese University of Hong Kong

14:50 Quantile Regression for Censored Data with a Cure Fraction
Bo Han
Yunnan University

S070: Modern Statistical Learning and High Dimensional Modeling

Room: Chi Wah Lecture Hall 204
Organizer: Xingqiu Zhao
Chair: Xingqiu Zhao

14:00 Acceleration for Diffusion Models: Theory and Practice
Gen Li
The Chinese University of Hong Kong

14:25 High-dimensional longitudinal model averaging via leave-subject-out cross-validation functionals
Shirong Deng
Wuhan University

14:50 A Conditional Distribution Equality Testing Framework Using Deep Generative Learning
Siming Zheng, ◆ Tong Wang, Meifang Lan, Yuanyuan Lin
Southeast University, Southeast University, The Chinese University of Hong Kong, The Chinese University of Hong Kong

15:15 Deep Nonparametric Inference for Conditional Mean Models with Panel Count Data
◆ *Anyin Feng, Qiang Wu, Xiangbin Hu and Xingqiu Zhao*
The Hong Kong Polytechnic University

S078: New Advances in Differential Privacy

Room: Chi Wah Lecture Hall 205
Organizer: Xuerong Chen
Chair: Yanqing Zhang

14:00 Truth Inflated Mechanism for Distributional Privacy
Yanqing Zhang
Yunnan University

14:25 Online Differentially Private Inference for Linear Regression Model
◆ *Senlin Yuan, Fang Liu, Xuerong Chen*
Yunnan University

14:50 Data Aggregation Based on Adaptive Personalized Local Differential Privacy
Peng Zhang
Kunming University of Science and Technology

15:15 A General Distributed Privacy-Preserving Statistical Inference Framework for Massive Data
Senlin Yuan, ◆ Xuerong Chen, Yong Zhou
Yunnan University, Southwestern University of Finance and Economics, East China Normal University

S082: New Developments of Statistical Methods in Bio-Medical Studies

Room: Chi Wah Lecture Hall 305
Organizer: Gengsheng Qin
Chair: Gengsheng Qin

14:00 Model Selection for Archimedean Copula Models Copula Models Based on Dependent Censored Data
Antai Wang
New Jersey Institute of Technology

14:25 Adaptive Metric Learning on SPD Manifolds via the Alpha–Procrustes Framework
◆ *Shao-Wei Huang, Jingyi Zheng, Huajun Huang*
Auburn University, Auburn University, Auburn University

14:50 Interval Estimation for the Area Under the ROC Curve with Verification Bias
Shirui Wang, Shuangfei Shi, ◆ Gengsheng Qin
Georgia State University, Georgia State University, Georgia State University

15:15 Latent Single-Index Models with Factor Structure for Multivariate Ordinal DataI
Zhiyong Chen
Fujian Normal University

S086: Novel Statistical and Machine Learning Methods for Analyzing Omics Data

Room: Chi Wah Lecture Hall 107
Organizer: Shili Lin
Chair: Hongmei Jiang

14:00 Contextual Evaluation of MicroRNA Sequencing Data Harmonization in Machine Learning Tasks
Li-Xuan Qin
Memorial Sloan Kettering Cancer Center

14:25 Exploration of Innovative Deep Learning Methods for Cancer Studies Based on Multi-Omics Data
◆ *Zheng Xu, Cong Wu, Yan Su, Changzhi Ma*

Wright State University, Wright State University, Medpace, College of Wooster

14:50 A Divide and Conquer Strategy for Recapitulating Whole Genome 3D Structure Using Hi-C Data
Jincheol Park, ⬩ *Meng Wang, Emerson Webb, Victor Jin, Shili Lin*
Abigail Wexner Research Institute at Nationwide Children's Hospital

15:15 Bayesian Methods for Phenotype-Associated Domain and Gene Detection in Spatial Transcriptomics
Han Wang
China Agricultural University

S099: Recent Advances in High-Dimensional and Functional Hypothesis Testing

Room: Chi Wah Lecture Hall 210
Organizer: Fengting Yi
Chair: Fengting Yi

14:00 A Unified Estimation and Inference Framework for High-Dimensional Time Series Models
Jinyuan Chang, ⬩ *Qiao Hu, Zhentao Shi, Jia Zhang*
Southwestern University of Finance and Economics, The Chinese University of Hong Kong, Chinese University of Hong Kong, Southwestern University of Finance and Economics

14:25 High-Dimensional Convoluted Rank Regression with Latent Factors
⬩ *Chenye Zhao, Yanmei Shi, Xu Guo*
Beijing Normal University, Beijing Normal University, Beijing Normal University

14:50 Goodness-of-Fit Test for Generalized Functional Linear Models via Projection Averaging
Feifei Chen, Kaiming Zhang, ⬩ *Yanni Zhang, Hua Liang*
Beijing Normal University, Beijing Normal University, Beijing Normal University, George Washington University

S100: Recent Advances in High-Dimensional Data Analysis

Room: Chi Wah Lecture Hall 306
Organizer: Chiung-Yu Huang
Chair: Ying Sheng

14:00 Efficient Machine Unlearning With Minimax Optimality
Jingyi Xie, Linjun Zhang, ⬩ *Sai Li*
Renmin University of China, Rutgers University, Tsinghua University

14:25 Model-X Knockoff Framework for Genome-Wide Survival Association Analysis
Shiyang Ma
Shanghai Jiao Tong University

14:50 Experimental Design and Differential Inference for Comparative Single-Cell RNA-Sequencing Studies
⬩ *Fangda Song, Kevin Y. Yip, Yingying Wei*
The Chinese University of Hong Kong, Shenzhen, Sanford Burnham Prebys Medical Discovery Institute, The Chinese University of Hong Kong

15:15 Information-Incorporated Estimation of High-Dimensional Generalized Linear DAGs with Interventions
⬩ *Weijuan Liang, Qingzhao Zhang, Shuangge Ma*
Xiamen University, Xiamen University, Yale School of Public Health

S115: Recent Advances on the Analysis of Failure Time Data

Room: Chi Wah Lecture Hall 307
Organizer: Jianguo Sun
Chair: Mingyue Du

14:00 Longitudinal First Hitting-Time Models for Survival Data
Mei-Ling Ting Lee
University of Maryland

14:25 Beyond Hazard-Based Illness-Death Models for Semi-Competing Risks Data
Rui Huang, ⬩ *Liming Xiang*
Nanyang Technological University, Nanyang Technological University

14:50 Prognostic Assessment of High Dimensional Survival Targets with Auxiliary Information Transfer
Jie Ding, Jialiang Li, ⬩ *Xiaoguang Wang*
Dalian Maritime University, National University of Singapore, Dalian University of Technology

15:15 An Optimal Two-Step Estimation Approach for Two-Phase Studies
⬩ *Qingning Zhou, Kin Yau Wong*
The University of North Carolina at Charlotte, The Hong Kong Polytechnic University

S121: Scalable and Distributed Statistical Learning for Complex Data Structures

Room: Chi Wah Lecture Hall 108
Organizer: Lu Tang
Chair: Lu Tang

14:20 Synthetic Data for Imbalanced Learning
⬩ *Anru Zhang, Zhengchi Ma*

Duke University, Duke University

14:25 Covariance Graph-Guided Harmonization in Distributed Neuroimaging Studies
Yiwen Dong, Jingsong Zhou, Shuo Chen, Dana Tudorascu, ⋆Qiong Wu
University of Pittsburgh, University of Maryland, University of Maryland, University of Pittsburgh, University of Pittsburgh

14:50 Bayesian Intelligence for Brain–Computer Interfaces
Jian Kang
University of Michigan

S135: Statistical Innovations for Spatial Transcriptomics and 3D Genomics

Room: Chi Wah Lecture Hall 201
Organizer: Li-Xuan Qin
Chair: Jian Zou

14:00 When Sparsity Meets Spatial Structure: Fusion of Sparsity and Spatial Information for Clustering Hi-C Single-Cell Data
Shili Lin
Ohio State University

14:25 Unveiling Fine-Scale Spatial Structures and Amplifying Gene Expression Signals in Ultra-Large ST Slices with HERGAST
Yuqiao Gong, Xin Yuan, ⋆Zhangsheng Yu
Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University

14:50 Statistical Methods for Cell-Cell Interaction Studies on Spatially Resolved Transcriptomics
Xiaoyu Song
National University of Singapore

15:15 Spatially Aware Adjusted Rand Index for Evaluating Spatial Transcriptomics Clustering
Yinqiao Yan, Xiangnan Feng, ⋆Xiangyu Luo
Beijing University of Technology, Fudan University, Renmin University of China

S139: Statistical Learning with AI

Room: Chi Wah Lecture Hall 206
Organizer: Jiancheng Jiang
Chair: Jiancheng Jiang

14:00 A Sparse Bayesian Learning Algorithm for Estimation of Interaction Kernels in Motsch-Tadmor Model
⋆Jinchao Feng, Sui Tang
Great Bay University, University of California, Santa Barbara

14:25 Score-based Generative Models with Gamma Processes
Libai Xu
Soochow University

14:50 Neural Dynamic Time Warping A Differentiable Framework for Financial Bubble Analysis
⋆Lukun Zheng, Yufei Zheng
Western Kentucky University, Shandong University

15:15 A Max 0-1 Weighted Log-Rank Test for Time-to-Event Data with Random Delayed Treatment Effect
⋆Jiarui Liu, Chenghao Chu, Bingming Yi, Ming-Hui Chen
Great Bay University

S146: Structured Data Analysis and Dependence Modeling

Room: Chi Wah Lecture Hall 207
Organizer: Fang Yao
Chair: Ying Yang

14:00 Flow-Driven Regression for Spatially Heterogeneous Coefficients
Wendong Li
East China Normal University

14:25 Deep Sparse Masks via Optimal Transport
Zihao Tang, ⋆Yixuan Qiu
Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

14:50 Dependence-Adaptive Modelling for Functional Time Series
⋆Zerui Guo, Jianbin Tan, Hui Huang
Sun Yat-sen University, Duke University, Renmin University of China

15:15 Deconfounding via Profiled Transfer Learning
⋆Ziyuan Chen, Yifan Jiang, Jingyuan Liu, Fang Yao
Peking University, The Pennsylvania State University, Xiamen University, Peking University

S150: The Frontier Research of Statistical Process Quality Control Under AI Era

Room: Chi Wah Lecture Hall 301
Organizer: Chunjie Wu
Chair: Chunjie Wu

08:30 A Change Point Control Chart Based on an EWMA-Type Smoothing of the Cross Matching A1 Statistics
Jiujun Zhang
Liaoning University

08:55 A Robust Online Control Chart Based on the Adaptive Conformal Method
⋆Miaomiao Yu, Yu Cao

East China Normal University

09:20 Online Monitoring in Telecommunication Signals via Wasserstein Barycenter Modeling
Wei Zhao
Shanghai Polytechnic University

09:45 A Bayesian Functional Tensor Framework for Online Monitoring of Continuous-Indexed Dynamic Processes with Application to Meteorological Surveillance
Yifan Li
Nanjing Audit University

S144: Statistics and Machine Learning at the Frontiers of Genomics and Genetics

Room: Chi Wah Lecture Hall 208
Organizer: Hongkai Ji
Chair: Hongkai Ji

14:00 Decoding and AI-Driven Design of Cell Type Specific Cis-Regulatory Elements
Yuan He
Beijing Zhongguancun Academy

14:25 Spatially Resolved Mapping of Cells Associated with Human Complex Traits
Jian Yang
Westlake University

14:50 AI for (Bio) Statistics and (Bio) Statistics for AI
Yiqun Chen
Johns Hopkins University

15:15 Bayesian Rhythmic Model for Jointly Detecting Circadian Biomarkers and Predicting Molecular Circadian Time in Human Post-Mortem Brain Transcriptome
George C. Tseng
University of Pittsburgh

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

S001: Advances in Statistical Methods for Genetic and Genomic Studies

Room: Chi Wah Lecture Hall 201
Organizer: Jin Liu
Chair: Jin Liu

16:00 Integrative Analysis and Regulatory Inference for Spatial Multi-Omics Data
Jishuai Miao, Jinzhao Li, Jingxue Xin, Jiajuan Tu, Muyang Ge, Ji Qi, Xiaocheng Zhou, Ying Zhu, Can Yang, ⋆ *Zhixiang Lin*
The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong, Hubei University of Technology, The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong, Fudan University, The Hong Kong University of Science and Technology, The Chinese University of Hong Kong

16:25 A Flexible and Unified Framework for Single- and Multi-Outcome Mendelian Randomization Using Summary Statistics
Lin Chen
University of Chicago

16:50 Deep Learning-Based Hierarchical Community Detection for High-Dimensional Gene Regulatory Networks
Jarred Kvamme, Jordan Davis, ⋆ *Audrey Fu*
University of Idaho, New Jersey Institute of Technology, Wayne State University

17:15 Building a Causality-Aware Single-Cell RNA-seq Foundation Model via Context-Specific Causal Regulation Modeling
Kaichen Xu, Yuxuan Shen, Maocheng Zhu, Zhenke Duan, Jin Liu, ⋆ *Xiaobo Sun*
Emory University, Georgia Institute of Technology, Zhongnan University of Economics and Laws, Zhongnan University of Economics and Laws, The Chinese University of Hong Kong, Shenzhen, Emory University

S003: Advanced Methods for Dependence, Signals and Spatio-Temporal Detection

Room: Chi Wah Lecture Hall 202
Organizer: Zheyang Wu
Chair: Wen Zhou

16:00 Double Validity of Robust Standard Errors under Complex Dependence
⋆ *Wen Zhou, Lihua Lei, Peng Ding, Zifeng Zhang*
New York University

16:25 Invariance Theory for Adaptive Signal Detection
⋆ *Mengjiao Tang, Yao Rong, Augusto Aubry, Antonio De Maio*
Yunnan University, Yunnan University, University of Naples Federico II, University of Naples Federico II

16:50 Likelihood Ratio Tests for Poisson Graph with Directed Acyclicity Constraints
Xin Liu
Shanghai University of Finance and Economics

17:15 Supervised Spatial Metric Learning for Clustering and Prediction
Xinyue Zhang, ⋆ *Hong Gu, Abdrew Irwin, Tobias Kenney*
Dalhousie University, Dalhousie University, Dalhousie University, Dalhousie University

S015: Advances in Predictive and Integrative Modeling for Complex Health Data

Room: Chi Wah Lecture Hall 211
Organizer: Ming Wang
Chair: Lijun Zhang

16:00 Building Trust in Clinical AI: An Explainable Decision Support System for Chronic Kidney Disease
Krishna Mridha, Ming Wang, ◆ *Lijun Zhang*
Case Western Reserve University, Case Western Reserve University, Case Western Reserve University

16:25 Variational Bayesian Inference for Poissonian Image Restoration
Jinling Liu, ◆ *Chao Wang, Youwei Wen*
Southern University of Science and Technology, Southern University of Science and Technology, Hunan Normal University

16:50 BioBMS: A Knowledge-Guided Dynamic Bayesian Screening and Dual-Prior Validation Framework for Robust Biomarker Discovery
He Mu
Xi'an Jiaotong-Liverpool University

S016: Advances in Statistical Inference for Complex and High-Dimensional Models

Room: Chi Wah Lecture Hall 102
Organizer: Fang Yao
Chair: Fang Yao

16:00 Integral Probability Metric-Guided CUSUM-Net for Nonparametric Changepoint Detection
Zhou Yu
East China Normal University

16:25 Neural Wasserstein Two-Sample Tests
◆ *Xiaoyu Hu, Zhenhua Lin*
Xi'an Jiaotong University, National University of Singapore

16:50 Structural Change Detection in Dynamic Systems
◆ *Wei Zhang, Fang Yao*
Fudan University, Peking University

17:15 Semiparametric M-estimation with Overparameterized Neural Networks
Shunxing Yan
Peking University

S026: AI-Driven Genomics and Precision Medicine: From Prediction to Personalization

Room: Chi Wah Lecture Hall 104
Organizer: Ching-Ti Liu
Chair: Xiaoyu Song

16:00 MethylFM: A DNA Methylation Foundation Model for Modeling Epigenomic Regulatory Dynamics
Limeng Pu, ◆ *Xiang Chen*
St. Jude Children's Research Hospital, St. Jude Children's Research Hospital

16:25 Toward Robust Genomic Prediction Across Asian Ethnicities: Insights from Machine Learning Approaches
◆ *Qiao Fan, Tingtong Liu, Xiaomin Liu, minhao Yao*
Duke-NUS Medical School, Duke-NUS Medical School, Duke-NUS Medical School, Duke-NUS Medical School

16:50 AdaBoost-rDNN: A Boosting Ensemble of Deep Neural Networks with Regularization for Disease Outcomes
◆ *Yaqing Xu, Mengni Zhang, Kuangnan Fang, Shuangge Ma*
Shanghai Jiao Tong University School of Medicine, Xiamen University, Xiamen University, Yale School of Public Health

S024: Advancing Complex Dependent Data Analysis

Room: Chi Wah Lecture Hall 301
Organizer: Fukang Zhu
Chair: Zhengwei Liu

16:00 Mixture of Dynamic Experts: Traffic Flow
◆ *Zhengwei Liu, Sheng Jiang*
Nanjing Normal University, The Chinese University of Hong Kong, Shenzhen

16:25 Robust Estimation and Bootstrap Testing for Self-Excited Threshold Poisson Autoregressive Model
Xinyang Wang
Liaoning University

16:50 Several Types of Binomial Autoregressive Models with Endpoint Inflations and Their Statistical Inference
◆ *Jie Zhang, Xiaojing Fan, Kai Yang, Yao Kang*
Changchun University of Technology, Changchun University of Technology, Changchun University of Technology, Xi'an Jiaotong University

17:15 A Composite Framework for Correcting Under-Reporting in Count Data with Dependence
Yixuan Fan
Tianjin University

S034: Conditional Modeling and Causality: Generative Models for Learning and Integration

Room: Chi Wah Lecture Hall 106
Organizer: Jian Huang
Chair: Jian Huang

16:00 Semi-Supervised Conditional Diffusion via Label Augmentation
Jin Su, Yuan Gao, Yong Zhou, Jian Huang
The Hong Kong Polytechnic University, Nankai University, East China Normal University, The Hong Kong Polytechnic University

16:25 Causal Effect Heterogeneity Estimation Using Summary Statistics
Xingjie Shi
East China Normal University

16:50 From Conditional to Unconditional Independence: Testing Conditional Independence via Transport Maps
Chenxuan He, Yuan Gao, Liping Zhu, Jian Huang
Renmin University of China, Nankai University, Renmin University, The Hong Kong Polytechnic University

17:15 Conditional Augmentation Enables Effective Use of Synthetic Data in Diffusion Models
Xueyu Zhou, Yuan Gao, Jian Huang
The Hong Kong Polytechnic University, Nankai University, The Hong Kong Polytechnic University

S047: Frontiers in Discrete-Valued Time Series Modeling

Room: Chi Wah Lecture Hall 303
Organizer: Fukang Zhu
Chair: Fukang Zhu

16:00 Z-Valued Smooth Transition GARCH Models: Specification and Testing
Fukang Zhu
Jilin University

16:25 Inference for Z-Valued INGARCH Models Allowing for Negative Autocorrelation
Yuqing Zhang, Baichuan Lin, Yao Kang
Xi'an Jiaotong University, Xi'an Jiaotong University, Xi'an Jiaotong University

16:50 Quantile Regression and Inference for Count Network Time Series
Danshu Sheng
Harbin Institute of Technology

17:15 Forecasting Natural Disaster Frequencies Using Nonstationary Count Time Series Models
Jian Pei, Yang Lu
Beijing University of Civil Engineering and Architecture, Concordia University

S048: Frontiers in Learning and Optimization: From GenAI to Dynamic Operations

Room: Chi Wah Lecture Hall 107
Organizer: Zhengling Qi
Chair: Zeyu Bian

16:00 **Robust Sequential Experimental Design for A/B Testing**
Qianglin Wen, Xiangkun Wu, Chengchun Shi, Ting Li, Niansheng Tang, Yingying Zhang, Hongtu Zhu
Yunnan University, Zhejiang University, East China Normal University

16:25 Statistical Inference in an Interactive Learning Paradigm
Yuchen Wu, Kangjie Zhou, Weijie Su
Cornell University, Columbia University, University of Pennsylvania

16:50 Integrating High-Dimensional Censored Data Under Privacy Constraints Via Localized Computations
Xin Ye, Yanyan Liu, Bingyao Huang
Guangdong University of Finance and Economics,Wuhan University, Guangdong University of Technology

S049: Frontiers in Methods for Complex-structured Survival Data

Room: Chi Wah Lecture Hall 103
Organizer: Huiqiong Li
Chair: Huiqiong Li

16:00 Distributed Privacy-Preserving Group Inference for High-Dimensional Generalized Linear Models
Yingrui Wang, Xuerong Chen, Liuquan Sun, Jian Huang, Min-Qian Liu, Dongxiao Han
Nankai University, Southwestern University of Finance and Economics, Chinese Academy of Sciences and School of Mathematical Sciences, The Hong Kong Polytechnic University, Nankai University, Nankai University

16:25 A Double-Semiparametric Approach for Extending Mixture Cure Models with Interval-Censored Data
Xiaoyu Liu
Jinan University

16:50 Dually Robust Nonparametric Censored Quantile Regression with Deep Neural Networks
Ziqiang Deng, Huiqiong Li, Niansheng Tang, Liuquan Sun
Yunnan University, Yunnan University, Yunnan University, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

17:15 Decoding AI Precision Medicine Paradox: A Fairness-Aware Optimal Transport
Zhimiao Cao
Shandong University

S060: Interval-Valued Data Modelling and Forecasting

Room: Chi Wah Lecture Hall 105

Organizer: Yuying Sun
Chair: Yuying Sun

16:00 Least Hausdorff Distance Regression for Interval-Valued Data
Wei Lin, Xiaojun Song
University of International Business and Economics, Peking University

16:25 Self-Exciting Hysteretic Interval Autoregressive Processes
Han Li, Xinyu Bi, Kai Yang, Shishun Zhao
Changchun University, Changchun University, Changchun University of Technology, Jilin University

16:50 Optimal Parameter-Transfer Learning for Interval-valued data
Haowen Bao, Yongmiao Hong, Yuying Sun, Shouyang Wang
Xidian University, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

17:15 Autoregressive Models for Functional Interval-Valued Time Series Data
jianhao Xu
University of Chinese Academy of Sciences

S066: Modern Methods for Omics Inference and Clinical Translation

Room: Chi Wah Lecture Hall 304
Organizer: Ming Wang
Chair: Weihong Guo

16:00 Unrolling Regularized Non-Negative Factorization for Spatial Transcriptomics Deconvolution
Weihong Guo, Liangliang Zhang, Ruitao Liu
Case Western Reserve University, Case Western Reserve University, University of Utah

16:25 CoReg: Co-Expression Network Multivariate Regression for Improved Replicability
Shuo Chen, Hwiyoung Lee
University of Maryland, University of Maryland

17:15 Bridging Data Gaps in Oncology: Large Language Models and Recommendations for Cancer Treatment in Clinical Trial Applications
Tengjie Tang, Angkai Li, Xuanyu Bai, Xingye Tan, Qingli Ji, Michael Ulis, Lu Si, Le Bao
Pennsylvania State University

S068: Modern Perspectives on High-Dimensional Statistical Theory and Methodology

Room: Chi Wah Lecture Hall 305
Organizer: Tianying Wang
Chair: Yuqian Zhang

16:00 Generalized Tensor Completion with Non-Random Missingness
Jingfei Zhang
Emory University

16:25 Facilitating Model-Based Clustering Analysis by Dimension Reduction
Yin Jin, Wei Luo
Zhejiang University, Zhejiang University

16:50 MintCNA: A Deep Learning Method for Integrative Copy Number Profiling with Single-Cell Multi-Omics Data
Wenhan Bao, Feifei Xiao
University of Florida, University of Florida

17:15 Semi-Supervised Linear Regression: Enhancing Efficiency and Robustness in High Dimensions
Kai Chen, Yuqian Zhang
Renmin University of China, Renmin University of China

S071: Modern Statistical Learning for Complex Data Analysis

Room: Chi Wah Lecture Hall 112
Organizer: Lexin Li
Chair: Ting Li

16:00 Active Hypothesis Testing under Computational Budgets with Applications to GWAS and LLM
Yin Xia
Fudan University

16:25 Logistic Regression and Classification with Covariates in Metric Spaces
Yinan Lin, Zhenhua Lin
Chongqing Normal University, National University of Singapore

16:50 EnsLoss: Stochastic Calibrated Loss Ensembles for Preventing Overfitting in Classification
Ben Dai
The Chinese University of Hong Kong

17:15 Weighted Youden Index Maximization
Jialiang Li
National University of Singapore

S084: New Methods for the Analysis of Complex Data

Room: Chi Wah Lecture Hall 203
Organizer: Mingyue Du
Chair: Mingyue Du

16:00 Variational Bayesian Estimation for Joint Models of Longitudinal and Interval-Censored Fairure Time Data with Random Change-Points
Fengting Yi
Yunnan University

16:25 A Longitudinal Multi-Domain Latent Structure Model for Domain-specific Disease Progression
Zijian Ye,Yuanjia Wang,Yu Gu, ⋆ *Kai Kang*
Sun Yat-sen University,Columbia University,The University of Hong Kong,Sun Yat-sen University

16:50 Transfer Learning Estimation of the Accelerated Failure Time Model Based on High-Dimensional Data
⋆ *Yichen Lou, Mingyue Du, Hui Zhao, Jianguo Sun*
Nanyang Technological University, Singapore, Jilin University, Zhongnan University of Economics and Law, Southern University of Science and Technology

S085: New Statistical and Coputational Methods for Large Data Analysis

Room: Chi Wah Lecture Hall 204
Organizer: Tingting Zhang
Chair: Ziqi Chen

16:00 Causal Learning with Invalid Instruments for High-dimensional Imaging Responses
Yiting Wang, Chunlin Li, ⋆ *Shan Yu, Lily Wang*
University of Virginia, University of Virginia, University of Virginia, George Mason University

16:25 High-Dimensional Dynamic Covariance Models with Random Forests
Shuguang Yu, Fan zhou, Yingjie zhang, ⋆ *ziqi chen, hongtu zhu*
East China Normal University

16:50 Detection of Medication Taking Using Wrist-Worn Commercially Available Wearable Device
⋆ *Quy Cao, Amy Laughlin*
University of Pennsylvania, University of Pennsylvania

17:15 Multiple Change Point Detection in Directed Temporal Networks with Non-Gaussian Components
⋆ *Sehun An, Ting Fung Ma, Shanghong Xie*
University of South Carolina., University of South Carolina, University of South Carolina

S096: Recent Advances in Computational Methods for Single-Cell and Spatial Omics

Room: Chi Wah Lecture Hall 205
Organizer: Ruibin Xi
Chair: Ruibin Xi

16:00 Intelligent Decoding and Generation of Spatial Biology
Shihua Zhang
The Academy of Mathematics and Systems Science, Chinese Academy of Sciences

16:25 Intelligent Analysis and Mining of Spatial Omics Data
Zhiyuan Yuan
Fudan University

16:50 Predicting Unseen Perturbation Effects Using Data Similarity and Global Structure
Jingshu Wang
University of Chicago

17:15 Spatial-Aware Detection of Copy Number Alterations from Spatial Transcriptomics Using SpaCNA
Zihui Zhang, Xiaochen Wang, Hong Xuan, Yan Xu, ⋆ *Zijie Jin, Ruibin Xi*
Peking University, Peking University, Peking University, Chinese Academy of Medical Sciences & Peking Union Medical College, Peking University, Peking University

S103: Recent Advances in Network Analysis

Room: Chi Wah Lecture Hall 210
Organizer: Wei Lan
Chair: Wei Lan

16:00 Differentially Private Estimation and Inference for Spatial Autoregressive Models
Ziyi Kong, Yuxin Tao, ⋆ *Danyang Huang, Zhanrui Cai*
Renmin University of China

16:25 Recover Common Components in Heterogeneous Chain-Linked Matrices
⋆ *Xinyan Fan, Hao Mei, Shuangge Ma*
Renmin University of China, Renmin University of China, Yale School of Public Health

16:50 Graph-Based Multisample Comparison with Application to Feature Selection for Multi-Category Responses
⋆ *Dan Pu, Haomin Shi, Wei Lan, Chih-Ling Tsai*
Southwestern University of Finance and Economics

S118: Recent Progress in Dependent Data Analysis

Room: Chi Wah Lecture Hall 306
Organizer: Tsung-I Lin
Chair: Wei-Ying Wu

16:00 Hidden-Factor Modeling of Climate-Driven Losses: An Illustration Using U.S. Crop Indemnities
⋆ *Kun Chen, Xingzuo He, Rui Huang, Zhiwei Tong*
Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, Nanjing University of Chinese Medicine, University of

Iowa

16:25 Multivariate Contaminated Normal Linear Mixed Models with Censored and Missing Outcomes Applied to Alzheimer's Disease Data
Wan-Lun Wang, Tsung-I Lin
Cheng Kung University, Chung Hsing University

16:50 Scalable Bayesian Spatiotemporal Tensor Modeling for Censored and Missing Areal Data
Luis Mauricio Castro Cepero
Pontificia Universidad Católica de Chile

17:15 Generative Bootstrap Sampler for Scalable Inference Under Dependent Data
Fanghua Chen, Ting Fung Ma, Ray Bai
University of South Carolina, University of South Carolina, George Mason University

S133: Statistical Inference for Complex Data

Room: Chi Wah Lecture Hall 108
Organizer: Jingyuan Liu
Chair: Jingyuan Liu

16:00 Is the F-Test Doubly Robust?
Lucy Xia, Oliver Feng, Yang Feng, Min Xu, Richard Samworth
The Hong Kong University of Science and Technology, London School of Economics and Political Science, New York University, Rutgers University, University of Cambridge

16:25 Efficient Computing for High-Dimensional Portfolio Optimization
Songshan Yang
Renmin University of China

16:50 Conditional Selective Inference for the Selected Groups in Panel Data
Chuang Wan, Jiajun Sun, Xingbai Xu
Jinan University, Xiamen University, Xiamen University

17:15 Fairness Auditing with General Protected Attribute
Junjie Huang, Jingyuan Liu, Xianli Zeng
Xiamen University

S141: Statistical Methods for Genetics and Genomics

Room: Chi Wah Lecture Hall 308
Organizer: Zheyang Wu
Chair: Zheyang Wu

16:00 GenoSim: Feasible Correlations of Discrete Random Variables and Genotype Data Simulation
Hong Zhang, Zheyang Wu
Pfizer, Worcester Polytechnic Institute

16:25 Mendelian Randomization Methods for Pleiotropic and Weak Effects and Their Applications
Miaoxin Li, Bin Tang
Sun Yat-Sen University, Sun Yat-Sen University

16:50 Spatially-Informed Reference-Free Deconvolution for Spatial Transcriptomics with SpatialCD
Phuong Vo, Yuehua Cui
Michigan State University, Michigan State University

17:15 TD-TMA: Integrative Analysis of Multiple Ligand-Receptor Families for Sample Clustering via Tensor Decomposition Based Trimmed Model Averaging
Nayang Shan, Lin Hou, Xiting Yan, Zuoheng Wang
Capital University of Economics and Business, Tsinghua University, Yale University, Yale University

S142: Statistical Methods for Reliable Machine Learning

Room: Chi Wah Lecture Hall 206
Organizer: Hongxin Wei
Chair: Hongxin Wei

16:00 Personalized Bayesian Federated Learning with Wasserstein Barycenter Aggregation
Feng Zhou
Renmin University of China

16:25 A Unified Newton Framework for Perturbation of Structured Matrix Estimators
Ziliang Zhong, Shuyang Ling
New York University Shanghai, New York University Shanghai

16:50 Toward Exact Convergence in Byzantine-Robust Decentralized Learning: A Statistical Identification Approach
Siyuan Zhang, Chengde Qian, Xin Liu, Changliang Zou
Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai Jiao Tong University, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Nankai University

17:15 False Discovery Rate Control over Data-driven Subgroups: A Unified Conformalized Approach
Yuyang Huo, Xiaoning Wang, Liuhua Peng, Changliang Zou
Nankai University, Nankai University, Melbourne university, Nankai University

S143: Statistical Modeling and Inference for Large Language Models

Room: Chi Wah Lecture Hall 207
Organizer: Weijie Su
Chair: Chendi Wang

16:00 Fundamental Trade-Offs in Multi-Bit Watermarking of Stochastic Processes
Haiyun He, Yepeng Liu, Zhuoer Shen, Ziqiao Wang, Yongyi Mao, Yuheng Bu
The Hong Kong University of Science and Technology (Guangzhou), University of California, Santa Barbara, University of California, Santa Barbara, Tongji University, University of Ottawa, University of California, Santa Barbara

16:25 Provable Training Data Identification for Large Language Models
Hongxin Wei
Southern University of Science and Technology

16:50 Bilevel ZOFO: Efficient LLM Fine-Tuning and Meta-Training
Peiran Yu
University of Texas Arlington

17:15 Optimal Watermark Localization in Mixed-Source Large Language Model Texts
Xiang Li
University of Pennsylvania

S149: The Broad Roles of Design in Modern Statistical Analysis

Room: Chi Wah Lecture Hall 302
Organizer: Boxiang Wang
Chair: Lyuou Zhang

16:00 From Subsampling to Causal Policy Learning: Efficient Experimentation with Rich Outcomes
Zhiheng Zhang
Shanghai University of Finance and Economics

16:25 Axis-Neighbor Geometry and Length-Scale Identifiability in Gaussian Process Models
Chenlu Shi, Lin Wang
New Jersey Institute of Technology, Purdue University

16:50 Weighted Confidence Set via Randomized Allocation
Lingyou Pang, Yunan Wu, Chenglong Ye
University of California Davis, Tsinghua University, University of Kentucky

17:15 Non-Smooth Optimization in Distributed Learning Environments
Joseph Casey, Boxiang Wang, Keren Li
University of Alabama at Birmingham, The University of Iowa, University of Alabama at Birmingham

S153: Uncertainty Quantification in Modern Statistics and Machine Learning

Room: Chi Wah Lecture Hall 208
Organizer: Chiung-Yu Huang
Chair: Huaqing Jin

16:00 Conformal Selective Prediction with General Risk Control
Tian Bai, Ying Jin
Stanford University, University of Pennsylvania

16:25 Seeing Through Correlations: Disentangled Feature Importance
Jin-Hong Du, Kathryn Roeder, Larry Wasserman
The University of Hong Kong, Carnegie Mellon University, Carnegie Mellon University

16:50 Parameter Inference via Nonlinear Partial Differential Equations Informed Gaussian Processes
Zhaohui Li, Shihao Yang, Jeff Wu
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Georgia Institute of Technology, The Chinese University of Hong Kong, Shenzhen

17:15 A Conformal Test of Linear Models via Permutation-Augmented Regressions
Leying Guan
Yale School of Public Health

S155: Matrix Data

Room: Chi Wah Lecture Hall 309
Organizer: Wang Zhou
Chair: Jiang Hu

16:00 High-Dimensional Spectral Enhancement Discriminant Analysis
Yonghan Zhang, Zhangni Pu, Lu Yan, Jiang Hu
Northeast Normal University, Northeast Normal University, Northeast Normal University, Northeast Normal University

16:25 The Spurious Factor Dilemma: Robust Inference in Heavy-Tailed Elliptical Factor Models
Jiang Hu, Jiahui Xie, Yangchun Zhang, Wang Zhou
Northeast Normal University, National University of Singapore, Shanghai University, National University of Singapore

16:50 On High-Dimensional Adaptive Sign Test
Chi Yao, Guangming Pan
Anhui University, Nanyang Technological University

17:15 Consistent and Powerful CUSUM Change-Point Test for Panel Data with Changes in Variance
Wenzhi Yang
Anhui University

S032: Causal Learning in Modern Health and Survey Data

Room: Chi Wah Lecture Hall 307
Organizer: Lu Tang
Chair: Lu Tang

16:00 Mediation Analysis for Complex Survey Designs
Wei Hao
University of Michigan

16:25 Debiased Machine Learning of Average Dose-Response Functions with General Instrumental Variables
Shuyuan Chen
Zhejiang University

16:50 Nonparametric Estimation of Subgroup Mediation Effects with Semi-Competing Risks Data
• *Zi Wang, Jialiang Li, Yu Cheng*
University of Electronic Science and Technology of China, National University of Singapore, University of Pittsburgh

17:15 A Copula-Based Model for Causal Effects with Interval-Censored Data under Semi-Competing Risks
• *Tao Sun, Xuezhu Qi*
Renmin University of China, Renmin University of China

e. June 29th AM (08:30-10:10)

S011: Advances in High-Dimensional Dependence Modeling

Room: Chi Wah Lecture Hall 115
Organizer: Ching-Kang Ing
Chair: Ching-Kang Ing

08:30 Structural Identification for Spatial-Temporal Dynamic Models
Rongmao Zhang
Zhejiang Gongshang University

08:55 Spatial Discrete Data Learning
Baisuo Jin
University of Science and Technology of China

09:20 Estimation and Inference for High-Dimensional Spatial Asset Pricing Models with Heterogeneous Interactions
• *Zhaoxing Gao, Sihan Tu, Ruey Tsay*
University of Electronic Science and Technology of China, Zhejiang University, University of Chicago

09:45 Covariate Shift Correction for Spatial Autoregressive Mode 1: A Reweighting Approach
• *Li Hou, Baisuo Jin, Ching-Kang Ing*
University of Science and Technology of China, University of Science and Technology of China, Tsing Hua University, Taiwan

S018: Advances in Statistical Machine Learning Methods for Complex Data

Room: Chi Wah Lecture Hall 106
Organizer: Changcheng Li
Chair: Changcheng Li

08:30 Functional Auxiliary Information Synthesis in Survival Analysis
• *Jie Ding, Jialiang Li, Ingrid Van Keilegom, Xiaoguang Wang*
Dalian Maritime University, National University of Singapore, Katholieke Universiteit Leuven, Dalian University of Technology

08:55 Fair Regression in Reproducing Kernel Hilbert Spaces: Single-Machine and Decentralized Implementations under Conditional Mean Parity
Xiaojun Mao
Shanghai Jiao Tong University

09:20 Integral Probability Metric-Guided CUSUM-Net for Nonparametric Changepoint Detection
Guanghui Wang
Nankai University

09:45 Variational Autoencoding Discrete Diffusion with Enhanced Dimensional Correlations Modeling
• *Tianyu Xie, Shuchen Xue, Zijin Feng, Tianyang Hu, Jiacheng Sun, Zhenguo Li, Cheng Zhang*
Peking University, Chinese Academy of Sciences, Huawei Foundation Model Department, The Chinese University of Hong Kong, Shenzhen, Huawei Foundation Model Department, Huawei Foundation Model Department, Peking University

S037: Decoding Complexity: Genomic Diversity, Disease Mechanisms, and Evolutionary Insights

Room: Chi Wah Lecture Hall 204
Organizer: Ching-Ti Liu
Chair: Zuoheng Wang

08:30 Radiogenomics of Smoking-Related Lung Disease: CT Radiomics Integrates with Airway Transcriptomics
Fenghai Duan
Brown University

08:55 Predicting Influenza Virus Evolution with Site-Based Mutation Dynamic Model
• *Maggie Haitian Wang, Chris Mok, Lirong Cao*
The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

09:20 Estimation of Direct and Indirect Polygenic Effects and Gene-Environment Interactions Using Polygenic Scores in Case-Parent Trio Studies
• *Ziqiao Wang, Nilanjan Chatterjee*
University of Virginia, Johns Hopkins University

09:45 Refining Structural Variation–Trait Associations with SVrefiner
Xiaoxuan Xia, Junteng Wu, Zhezhi Gao, Peixiong Yuan, Sheng Ye
Tianjin University, Tianjin University, Tianjin University, Tianjin University, Tianjin University

S043: From Data to Biomarkers: Statistical Innovation for Complex Challenges in Modern Genetics, Genomics and Proteomics

Room: Chi Wah Lecture Hall 107
Organizer: Li-Xuan Qin
Chair: Li-Xuan Qin

08:30 Horizontal Pleiotropy in Mendelian Randomization: Use It If You Cannot Beat It!
Anagh Chattopadhyay, Nilanjan Chatterjee
Johns Hopkins University

08:55 Statistical Tools for Integrative Proteogenomic Analysis
Pei Wang
Icahn School of Medicine at Mount Sinai

09:20 Quantile-Based Discriminant Analysis for High-Dimensional Imbalanced Classification
Qian Tang, Yuwen Gu, Boxiang Wang
University of Minnesota, University of Connecticut, University of Iowa

09:45 Using Multiple Biomarkers for Patient Enrichment in Two-Stage Clinical Designs
Fengqing Zhang
Drexel University

S044: From Probabilistic Theory to Causal Inference in Complex Systems

Room: Chi Wah Lecture Hall 205
Organizer: Doudou Zhou
Chair: Shuting Shen

08:30 Anti-Concentration Inequalities for the Difference of Maxima of Gaussian Random Vectors
Alexandre Belloni, Ethan Fang, Shuting Shen
Duke University, Duke University, National University of Singapore

08:55 Estimating Treatment Effect Sizes Without Knowing Who Was Treated
Yao Zhang
National University of Singapore

09:45 Mediation Analysis with Interval-Censored Outcome and Interval-Censored Mediator
Dongdong Li, Rui Wang
Harvard University, Harvard University

S051: High Dimensional Data Analysis and Change Point Inference

Room: Chi Wah Lecture Hall 201
Organizer: Long Feng
Chair: Long Feng

08:30 Two-Way Node Popularity Model for Directed and Bipartite Networks
Jiangzhou Wang
Shenzhen University

08:55 Adaptive Change Point Inference for High Dimensional Time Series with Temporal Dependence
Xiaoyi Wang, Jixuan Liu, Long Feng
Beijing Normal University, Nankai University, Nankai University

09:20 Simultaneous Detection and Localization of Mean and Covariance Changes in High Dimensions
Junfeng Cui, Guangming Pan, Guanghui Wang, Changliang Zou
Shenzhen University, Nanyang Technological University, Nankai University, Nankai University

09:45 Adaptive L-Statistics for High Dimensional Test Problem
Huifang Ma, Long Feng, Zhaojun Wang
Nankai University, Nankai University, Nankai University

S064: Model Averaging and Related Topics

Room: Chi Wah Lecture Hall 108
Organizer: Xinyu Zhang
Chair: Huihang Liu

08:30 Mallows-Type Model Averaging: Non-Asymptotic Analysis and All-Subset Combination
Jingfu Peng
Tsinghua University

08:55 A Scalable Model Averaging Based on Kullback-Leibler Distance for Multivariate Regression Models
Jie Zeng
Hefei Normal University

09:20 Handling Covariate Shift by Model Averaging
Yifan Zhang, Xinyu Zhang
Beijing University of Technology, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

09:45 Trans-MA: Sufficiency-Principled Transfer Learning via Model Averaging
Xiyuan Zhang, Huihang Liu, Xinyu Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Shanghai University of Finance and

Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

S065: Model Averaging for Time Series Data

Room: Chi Wah Lecture Hall 210
Organizer: Yuying Sun
Chair: Yuying Sun

08:30 Model Averaging for Functional Quantile Regression
Wenchao Xu, Xinyu Zhang, Jeng-Min Chiou, Yuying Sun
Shanghai University of International Business and Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Taiwan University & Academia Sinica, Taiwan, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

08:55 Model Uncertainty and Averaging of Spatial Autoregression for Spatial Heterogeneity
Yuge Luo, Wangli Xu, Zudi Lu, Jun Liao
Renmin University of China, Renmin University of China, City University of Hong Kong, Renmin University of China

09:20 Data Decomposition-Based Model Averaging
Yongmiao Hong, Wencan Lin, Yuying Sun, Shouyang Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

09:45 Optimal Parameter-Transfer Learning by Time-Varying Model Averaging
Boyan Li, Yuying Sun, Xinyu Zhang
Chinese Academy of Sciences, Chinese Academy of Sciences, Chinese Academy of Sciences

S072: Modern Statistical Methods and Artificial Intelligence: From Theory to Biomedical Applications

Room: Chi Wah Lecture Hall 302
Organizer: Ling Zhou
Chair: Ling Zhou

08:30 The Future of AI Scientists: Emerging Directions and Fundamental Challenges
Jie Ding
University of Minnesota

08:55 Mismatch-Aware Transfer Learning for Optimal Linear Prediction
Jinhong Li, Lu Tang, Guan Yu
University of Pittsburgh, University of Pittsburgh, University of Pittsburgh

09:20 Hierarchical Contrastive Learning for Multimodal Data with Partial Sharing
Huichao Li, Junhan Yu, Doudou Zhou
National University of Singapore

09:45 High-Dimensional Covariate-Augmented Overdispersed Poisson Factor Model with Application to Single-Cell Multi-Omics
Wei Liu, Qingzhi Zhong
Sichuan University, Jinan University

S077: Mordern Distributional Methods for Learning, Inference, and Robustness

Room: Chi Wah Lecture Hall 202
Organizer: Jian Huang
Chair: Jian Huang

08:30 Conditional Distribution Testing via Measure Transport
Yuan Gao, Chenxuan He, Jian Huang, Liping Zhu
Nankai University, Renmin University of China, The Hong Kong Polytechnic University, Renmin University of China

08:55 Transfer Learning through Enhanced Sufficient Representation: Enriching Source Domain Knowledge with Target Data
Yeheng Ge, Xueyu Zhou, Jian Huang
The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

09:20 Wasserstein Distributionally Robust Nonparametric Regression
Changyu Liu, Yuling Jiao, Junhui Wang, Jian Huang
Wuhan University, Wuhan University, The Chinese University of Hong Kong, The Hong Kong Polytechnic University

09:45 A Two-Armed Bandit Approach for Bang-Bang Optimal Stochastic Control with Generalized Cost Functionals
Zengjing Chen, Jianqi Feng, Lu Wang, Chengzhi Xing
Shandong University, Shandong University, The Hong Kong Polytechnic University, Zhongtai Securities

S081: New Developments in Survey Sampling

Room: Chi Wah Lecture Hall 307
Organizer: Guohua Zou
Chair: Wangxue Chen

08:30 Imperfect Extreme Ranked Set Sampling Design: A Framework for Robust and Efficient Inference
Shuangxi Liu, Wangxue Chen, A. M. Elsawah

Jishou University, Jishou University, Beijing Normal-Hong Kong Baptist University

08:55 An improved AdRec model for Missing data in Population Census
Lingli Sun, Tingting Zhang, Wenqiang Du, Shanshan Qin
Tianjin University of Finance and Economics

09:20 Measuring Cheating and Noncompliance in Survey with Sensitive Questions
Qin Wu
South China Normal University

09:45 Jackknife Model Averaging for Complex Survey Data
Huanxue Zhuang, Xianpeng Zong
Beijing University of Technology, Beijing University of Technology

S083: New Frontiers in Spatial Modeling: Dissimilarity, Extremes, and High-Dimensional Clustering

Room: Chi Wah Lecture Hall 206
Organizer: Kun Chen
Chair: Likun Zhang

08:30 Integrating Mechanistic Dynamics to Characterize the Causes of Extremes in Environmental Processes
Megha Pailan, Likun Zhang, Christopher K. Wikle
University of Missouri - Columbia, University of Missouri - Columbia, University of Missouri - Columbia

08:55 Generalized Chi-Squared Process for Spatial Dissimilarity
Xiaotian Zheng, Andrew Zammit-Mangion, Noel Cressie
University of Georgia, University of New South Wales, University of Wollongong

09:20 Cluster Size Distributions of Discrete Random Fields
Dan Cheng, John Ginos
Arizona State University, Arizona State University

09:45 BNPSpace: Bayesian Nonparametric Clustering for Spatially Resolved Transcriptomic Data
Guanyu Hu
Michigan State University

S087: Post and Semi-Supervised Inference

Room: Chi Wah Lecture Hall 211
Organizer: Qizhai Li
Chair: Haojie Ren

08:30 Algorithm-Agnostic Post-Clustering Differential Expression Analysis with Asymptotic FDR Control
Jiajun Liu, Mengtao Wen, Changliang Zou, Liu Liu
Nankai University, Nankai University, Nankai University, Chengdu University of Technology

08:55 Shape-Adaptive Conditional Calibration for Conformal Prediction via Minimax Optimization
Yajie Bao, Chuchen Zhang, Zhaojun Wang, Haojie Ren, Changliang Zou
Nankai University, Nankai University, Nankai University, Shanghai Jiao Tong University, Nankai University

09:20 Design and Modeling for Order-of-Addition Factorial Experiments with Interaction Effects
Jianbin Chen
Beijing Institute of Technology

09:45 Semi-Supervised Inference for U-Estimation
Liuhua Peng
University of Melbourne

S090: Recent Advance in Statistical Learning

Room: Chi Wah Lecture Hall 212
Organizer: Biao Cai
Chair: Biao Cai

08:30 False Discovery Rate Control via Data Splitting for Testing-after-Clustering
Lijun Wang, Yingxin Lin, Hongyu Zhao
Zhejiang University, The Chinese University of Hong Kong, Yale University

08:55 Robust Bayesian Inference on Riemannian Submanifold
Rong Tang, Anirban Bhattacharya, Debdeep Pati, Yun Yang
Hong Kong University of Science and Technology

09:20 Estimation and Inference for Nonparametric Expected Shortfall Regression over RKHS
Yue Wang
University of Science and Technology of China

S091: Recent Advances for Statistical Analysis of Censored Data

Room: Chi Wah Lecture Hall 303
Organizer: Hui Zhao
Chair: Hui Zhao

08:30 Transfer Learning for High-Dimensional Semiparametric Regression with Censored Covariates
Huiqiong Li
Yunnan University

08:55 Variable Selection for Interval-Censored Data with Informative Observation Times
Ni Li, Mengxin Wang
Hainan Normal University, Hainan Normal University

09:20 Weighted Jackknife Model Averaging Method for Right Censored Data Under Varying Coefficient Models
Binxia Liu
Taiyuan Normal University

09:45 Distributed Sparse Learning for Cox Models Over Networks: A Decentralized Approach
• *Yu Zhang, Hui Zhao*
Zhongnan University of Economics and Law, Zhongnan University of Economics and Law

S093: Recent Advances in Complex Data Analysis and New Applications

Room: Chi Wah Lecture Hall 308
Organizer: Yichuan Zhao
Chair: Lei Huang

08:30 Slacked Empirical Likelihoods for Post-Criterion Inference
Yiyuan She
Westlake University

08:55 Geometric standardized mean difference and its application to meta-analysis
Jiandong Shi, Xiaochen Zhang, Lu Lin, Hiu Yee Kwan, • *Tiejun Tong*
Chinese University of Hong Kong, Beijing Normal University at Zhuhai, Shandong University, Hong Kong Baptist University, Hong Kong Baptist University

09:20 Time-Varying High Quantile Estimation for Nonstationary Tail Dependent Time Series
• *Ting Zhang, Yu Shao*
University of Georgia, Ernst & Young

09:45 Model Estimation and Variable Selection for Longitudinal Partial Linear EV Models with Measurement Errors in All Covariates
Mingtao Zhao
Anhui University of Finance and Economics

S098: Recent Advances in Designs and Statistical Methods for Clinical Trials Research

Room: Chi Wah Lecture Hall 309
Organizer: Yuan Wu
Chair: Xiaofei Wang

08:30 Statistical Inference for Heterogeneous Treatment Effect with Right-Censored Data from Synthesizing Randomized Clinical Trials and Real-World Data
• *Guangcai Mao, Shu Yang, Xiaofei Wang*
Duke University, North Carolina State University, Duke University

08:55 Win Ratio Under Bivariate Mixed Censored Data Based on Death Time and Progression Time
• *Yuan Wu, Xiaofei Wang, Zhiguo Li*
Duke University, Duke University, Duke University

09:20 Deconfounded-Debiased Distributed Estimation for High-Dimensional Multi-Site Data with Heterogeneous Pervasive Hidden Confounders
Guoyou Qin
Fudan University

S105: Recent Advances in Selective Inference, Data Assimilation, and Functional Data

Room: Chi Wah Lecture Hall 304
Organizer: Anru Zhang
Chair: Wanjie Wang

08:30 Functional Regression on Product Manifolds with Application to Continuous Brain Connectivity
Yezi Wang, • *Lu Wang*
Central South University, Central South University

08:55 Localization Estimator for High Dimensional Tensor Covariance Matrices
• *Haoxuan Sun, Song Xi Chen, Yumou Qiu*
Harbin Institute of Technology

09:20 A Unified Framework for Large-Scale Inference of Classification: Error Rate Control and Optimality
• *Yinrui Sun, Yin Xia*
Fudan University, Fudan University

09:45 A New Approach to Conformalized Model Selection for Out-of-Distribution Testing
Rongyi Sun, Wenguang Sun, • *Zinan Zhao*
Zhejiang University, Zhejiang University, Zhejiang University

S111: Recent Advances in Statistical Methods for Genetic Data

Room: Chi Wah Lecture Hall 207
Organizer: Yingying Wei
Chair: Yingying Wei

08:30 Robust Multi-Ancestry Cis-Mendelian Randomization Method with Bayesian Fine-Mapping
Chengli Zhang, Chong Wu, • *Haoran Xue*
City University of Hong Kong, University of Texas MD Anderson Cancer Center, City University of Hong Kong

08:55 Knockoff-Augmented Meta-Analysis Identifies Shared and Ancestry-Specific Risk Loci in Multi-Ancestry Genome-Wide Association Studies
Jian Yin, • *Yi Yang*
City University of Hong Kong, City University of Hong

Kong

09:20 A Robust and Computationally Scalable Method for Integrating a Large Number of Genomic Studies
Zhiang Yuan, Hongyuan Cao, Yingying Wei
The Chinese University of Hong Kong, Florida State University, The Chinese University of Hong Kong

S119: Regression Modeling and Efficient Estimation for Complex Data

Room: Chi Wah Lecture Hall 305
Organizer: Liuquan Sun
Chair: Meiling Hao

08:30 Bayesian Spatiotemporal Modeling for the Inpatient Hospital Costs of Alcohol-Related Disorders
Zhen Yu
University of International Business and Economics

08:55 Subgroup Identification with Latent Factor Structure
Yong He, Dong Liu, Fuxin Wang, Mingjuan Zhang, Wen-Xin Zhou
Shandong University, National University of Singapore, University of Wisconsin-Madison, Shanghai Lixin University of Accounting and Finance, University of Illinois Chicago

09:20 A Nested Singular Value Decomposition-Based Surrogate Model for Spatiotemporal Flows
Shixiang Li
ZJUI Institute, Zhejiang University

09:45 Collaborative Estimation of Quantile Treatment Effect for Distributed Data
Ye Fan, Ying Wei, Sung Nok Chiu, Tiejun Tong, Nan Lin
Capital University of Economics and Business, Columbia University, Hong Kong Baptist University, Hong Kong Baptist University, Washington University in St. Louis

S132: Statistical Inference and Subgroup Analysis for Specialized Data Types

Room: Chi Wah Lecture Hall 306
Organizer: Liuquan Sun
Chair: Ye Fan

08:30 The MLE for Sparse Rasch Model: Consistency, Top-k Recovery and Gaussian Approximation
Pai Peng, Lianqiang Qu, Qiuping Wang, Shufang Wang, Ting Yan
Jianghan University, Central China Normal University, Zhaoqing University, Central China Normal University, Central China Normal University

08:55 Conditional Inference for High-Dimensional Multi-Omics Survival Data
Heyuan Zhang, Meiling Hao, Lianqiang Qu, Liuquan Sun
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, University of International Business and Economics, Central China Normal University, University of Chinese Academy of Sciences

09:20 Large-Scale Quantile Regression with Covariates Missing at Random
Xiaoran Yang, Yiwei Wang, Fangfang Bai
Shandong Technology and Business University, University of International Business and Economics, University of International Business and Economics

09:45 Online Inference Based on Streaming Time-to-Event Data
Xin Chen, Hua Liu, Jiaqi Men, Jinhong You
Shanghai Lixin University of Accounting and Finance

S137: Statistical Learning for Complex Data

Room: Chi Wah Lecture Hall 310
Organizer: Huiqiong Li
Chair: Huiqiong Li

08:30 Network Alternating Direction Method of Multipliers for Ultrahigh-Dimensional Decentralised Federated Learning
Wei Dong
Zhengzhou University

08:55 Transfer Learning for High-dimensional Accelerated Failure Time Models
Jie Wu
Anhui University

09:20 Minimum Distance Kernel Estimation Based Robust Perturbation Subsampling for Big Data Linear Regression
Mingqiu Wang
Qufu Normal University

09:45 A Novel Skinny Variational Bayesian Estimation Method for High-Dimensional Cox Models with Interval-Censored Data
Wenting Liu, Huiqiong Li, Niansheng Tang
Xi'an University of Finance and Economics, Yunnan University, Yunnan University

S145: Statistics, Machine Learning, and AI

Room: Chi Wah Lecture Hall 208
Organizer: Lexin Li
Chair: Ben Dai

08:30 Nested Deep Learning Model Towards a Foundation Model for Brain Signal Data
Haipeng Shen

University of Hong Kong

08:55 DeepSuM: A Deep Sufficient and Efficient Modality Learning Framework
Ting Li
Southern University of Science and Technology

09:20 In-Sample Evaluation of Subgroups Identified by Generic Machine Learning
Xinzhou Guo
Hong Kong University of Science and Technology

09:45 Signed Network Embedding: Balance, Community and Heterogeneity
Junhui Wang
Chinese University of Hong Kong

S057: Innovative Machine Learning and Statistical Methods for Complex Biomedical Data

Room: Chi Wah Lecture Hall 203
Organizer: Yuanjia Wang
Chair: Kai Kang

08:30 Multi-Modal Representation Learning with Partially Pairwise Observations
Younggeun Kim
Ulsan National Institute of Science and Technology

08:55 Evaluating Multiplex Diagnostic Test Using Multivariable Isotonic Regression
• *Ying Kuen Cheung, Louise Kuhn*
Columbia University, Columbia University

09:20 MACHINE: A Robust Multi-Ancestry Fine-Mapping Method Using a Continuous Shrinkage Prior
Xiang Li, Zewei Xiong, Pak Chung Sham, • *Yan Dora Zhang*
The University of Hong Kong, The University of Hong Kong, The University of Hong Kong, The University of Hong Kong

09:45 Dynamic Progression Model with Individualized Longitudinal Registration
• *Zexi Cai, Yuanjia Wang*
University of Macao

S151: Time Series Econometrics and Applications

Room: Chi Wah Lecture Hall 301
Organizer: Jiancheng Jiang
Chair: Shaoxin Hong

08:30 Robust Nonparametric Testing for Structural Changes in Multivariate Volatility via Multiple Quantiles
Jilin Wu
Xiamen University

08:55 A Testing for Symmetry via Quantile Regression Martingalization
• *Jinfeng Zhang, Shaoxin Hong*
Shandong university, Boston college

09:20 Estimating Temperature-Electricity Response Curves: A Mixed-Frequency Semiparametric Panel Data Model
Kerui Du
Xiamen University

09:45 Estimating the Number of Groups in Nonstationary Panels: A Penalty-Free Perspective
Ruiqi Liu, • *Shaoxin Hong*
Texas Tech University, Xiamen University

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

S004: Advanced Statistical and Machine Learning Methods in Genetics and Genomics

Room: Chi Wah Lecture Hall 106
Organizer: Tianying Wang
Chair: Bo Yu

10:30 A Scalable Approach for Dynamic Single-Cell eQTL Analysis
Lin Hou
Tsinghua University

10:55 A Powerful Transformation of Quantitative Responses for Biobank-Scale Association Studies
• *Yaowu Liu, Tianying Wang*
Southwestern University of Finance and Economics, Colorado State University

11:20 Rare Coding and Noncoding Variants Map 1,342 Diseases and Biomarkers in 490,549 Whole Genomes
Zilin Li
Northeast Normal University

11:45 CausalGRN Maps Causal Gene Regulatory Networks and Predicts Unseen Perturbation Effects from Single-Cell CRISPR Screens
• *Bo Yu, Dingyu Liu, Guanghao Qi, Danwei Huangfu, Li Hsu, Ali Shojaie, Wei Sun*
Fred Hutchinson Cancer Center, Sloan Kettering Institute, University of Washington, Sloan Kettering Institute, Fred Hutchinson Cancer Center, University of Washington, Fred Hutchinson Cancer Center

S008: Advances in Bayesian Modeling for Big and Complex Data

Room: Chi Wah Lecture Hall 204
Organizer: Tianzhou Ma
Chair: Hongjie Ke

10:30 A Semi-Reference Bayesian Approach to Cell-Type Annotation and Marker Discovery in Single-Cell RNA-seq
An-Shun Tai, Ying-Ting Huang
Tsing Hua University, Taiwan, Cheng Kung University

10:55 Bayesian Sparse Kronecker product decomposition
Shao-Hsuan Wang, Hsin-Hsiung 'Bill' Huang
Graduate Institute of Statistics of Central University, Taiwan, University of Central Florida

11:20 Synthesizing Multiple-Type Recurrent Events with the Terminal Event: A Joint Bayesian Copula-Frailty Modelling Approach
Menglu Liang, Yifan Yang, Lijun Zhang, Chixiang Chen, Ming Wang
University of Maryland, College Park, Case Western Reserve University School of Medicine, Case Western Reserve University School of Medicine, University of Maryland, Baltimore, Case Western Reserve University School of Medicine

S009: Advances in High-Dimensional and Complex Data Analysis

Room: Chi Wah Lecture Hall 205
Organizer: Changcheng Li
Chair: Changcheng Li

10:30 The Fairness Hierarchy: A Viewpoint from Causal Inference
Chengbo Zhang, Zhen Yao, Hao Pang, Changcheng Li
Dalian University of Technology, Dalian University of Technology, Dalian University of Technology, Dalian University of Technology

10:55 Pursuing Homogeneity and Sparsity in Simultaneous Quantile Regression
Zhen Zeng, Weixin Yao
Nanjing University of Finance and Economics, University of California, Riverside

11:20 Conditional Expectile-Based Value-at-Risk: Semiparametric Estimation and Inference
Feipeng Zhang
School of Economics and Finance, Xi'an Jiaotong University

11:45 Posterior Predictive (PoP) Design for Next-generation Phase I Clinical Trials
Yeqing Zhou
Tongji University

S017: Advances in Statistical Learning and AI for Biological and Health Sciences

Room: Chi Wah Lecture Hall 107
Organizer: Jun Li
Chair: Jun Li

10:30 Single-Cell Gene Regulatory Network Analysis with Applications to Alzheimer's Disease
Danni Liu, Zhongli Jiang, Yan Li, Dabao Zhang, Min Zhang
University of California, Irvine, University of California, Irvine, University of California, Irvine, University of California, Irvine, University of California, Irvine

10:55 Accelerating with AI: Reflections from a Computer Vision Project in Grass Seed Analysis
Yanming Di
Oregon State University

11:20 How Much Biology Do LLM-Derived Gene Embeddings Actually Know?
Yufei Cai, Dailin Gan, Hui Zhang, Jun Li
University of Notre Dame, University of Notre Dame, Northwestern University, University of Notre Dame

S020: Advances in Statistical Methods for Omics and Health Data Analysis

Room: Chi Wah Lecture Hall 108
Organizer: Tianzhou Ma
Chair: Menglu Liang

10:30 A Computational Pipeline for Studying Multicellular Interactions in 3D Bioprint Systems
Min Yang, Sha Cao
Oregon Health & Science University, Oregon Health & Science University

10:55 Advancing Computational Resources and Tools to Study Metabolism in Diseases
Chi Zhang, Sha Cao, Pengtao Dang, Paveethran Swaminathan
Oregon Health & Science University, Oregon Health & Science University, Oregon Health & Science University, Oregon Health & Science University

11:20 Correcting Bias from Covariate-Dependent Censoring in Survival Disparity Decomposition
Yan Li, Chong Zhang, Menglu Liang
University of Maryland, University of Maryland, University of Maryland

11:45 Inferring Non-Coding RNA Regulatory Network from Transcriptomic Data and Curated Databases
Hongjie Ke
Capital University of Economics and Business

S027: Applications in Biomedical Sciences

Room: Chi Wah Lecture Hall 305
Organizer: Grace S. Shieh
Chair: Hao Ho

10:30 Rank-Based Measures of Liquid Associatio
Hao Ho
University of California, Los Angeles

10:55 CRISP: A Modular Platform for Cryo-EM Image Segmentation and Processing
Szu-Chi Chung
Sun Yat-Sen University, Taiwan

11:20 Inference for Heavy-Tailed Integer-Valued Time Series with Applications to Infectious Disease Case Counts
Yuichi Goto, ⋆ *Gaspard Bernard*
Kyushu University, Academia Sinica, Taiwan

11:45 Toward Robust Machine Learning Methods for Individualized Treatment Effect Estimation
Keng-Te Liao
Tsing-Hua University, Taiwan

S046: Frontiers in Causal Inference: Addressing Time-Dependence, Fairness, and Long-Term Effects in Complex Data

Room: Chi Wah Lecture Hall 210
Organizer: Lu Wang
Chair: Lu Wang

10:30 Ensuring Fairness in Sequential Decision Making: Counterfactual Fairness in Reinforcement Learning for Opioid Use Disorder Services
Zhenke Wu
University of Michigan

10:55 On Robust Covariate Adjustment Methods in RCT
Lin Liu
Shanghai Jiao Tong University

11:20 Causal Mediation Analysis in Multistate Process
Yiqing Chen
Texas A&M University

11:45 A General Framework for Incorporating Identification Uncertainty in Individualized Treatment Rules
Muxuan Liang, Ting Ye, ⋆ *Yingqi Zhao*
MD Anderson Cancer Center, University of Washington, Fred Hutchinson Cancer Center

S053: High-Dimensional Statistics

Room: Chi Wah Lecture Hall 202
Organizer: Shurong Zheng
Chair: Xiaoyi Wang

10:30 A Non-Universal Convex Non-Gaussian Min–Max Theorem and Its Implications for High-Dimensional Machine Learning
Mohamed Chiheb Yaakoubi, Cosme Louart, Malik Tiomoko, ⋆ *Zhenyu Liao*
Chinese University of Hong Kong, Shenzhen, Chinese University of Hong Kong, Shenzhen, Huawei, Huazhong University of Science and Technology

10:55 Tensor Elliptical Graphical Model
Jixuan Liu, Zhengke Lu, ⋆ *Le Zhou, Long Feng, Zhaojun Wang*
Nankai University, Nankai University, Hong Kong Baptist University, Nankai University, Nankai University

11:20 Adaptive Detection of Change-Points for High-Dimension al Covariance Matrices
⋆ *Jiayu Lai, Xiaoyi Wang, Le Zhou, Shurong Zheng*
Jiangxi University of Science and Technology, Beijing Normal University, Hong Kong Baptist University, Northeast Normal University

11:45 Spectral Properties of High-dimensional Rescaled Sample Correlation Matrices under Elliptical Distributions
Weijiang Chen, Shurong Zheng, ⋆ *Tingting Zou*
Jilin University, Northeast Normal University, Jilin University

S067: Modern Mixture Model Methodology: Advances in Modeling, Inference, and Applications

Room: Chi Wah Lecture Hall 306
Organizer: Pengfei Li
Chair: Chunlin Wang

10:30 Gaussian Mixture Reduction Approach to Compacting 3D Gaussian Splatting in Computer Graphics
Tao Wang, Mengyu Li, Geduo Zeng, Cheng Meng, ⋆ *Qiong Zhang*
Renmin University of China, Tsinghua University, Renmin University of China, Renmin University of China, Renmin University of China

10:55 Model Positive and Unlabeled Data with a Generalized Additive Density Ratio Model
⋆ *Peijun Sang, Yifan Sun, Qinglong Tian, Donglin Zeng, Pengfei Li*
University of Waterloo, University of Waterloo, University of Waterloo, University of Michigan, University of Waterloo

11:20 Smoothed Likelihood for Combining Diagnostic Tests
Fangyong Zheng, Pengfei Li, ⋆ *Tao Yu*
National University of Singapore

11:45 Likelihood-Based Non-Parametric Receiver Operating Characteristic Curve Analysis in the Presence of Imperfect Reference Standard
Yifan Sun, Peijun Sang, Qinglong Tian, ◆Pengfei Li
University of Waterloo, University of Waterloo, University of Waterloo, University of Waterloo

S080: New Advances in Microbiome Data Analysis and Related Fields

Room: Chi Wah Lecture Hall 302
Organizer: Zhigang Li
Chair: Lihui Zhao

10:30 CAFT: A Compositional Log-Linear Model for Microbiome Data with Zero Cells
Mo Li, Glen Satten, ◆Ni Zhao
University of Louisiana at Lafayette, Emory University, Johns Hopkins University

10:55 Mediation Analysis with Compositional Exposures
Jing Ma
Fred Hutchinson cancer center

11:20 Joint Representation Learning, Feature Selection, and Clustering for Sparse Count Data with Applications to Single-Cell Genomics
Cheng Zhong,Siqi Jiang,◆Zhi Wei
New Jersey Institute of Technology, New Jersey Institute of Technology, New Jersey Institute of Technology

11:45 Estimating Equations with Inverse Probability Weighting for Microbiome Analysis
Zhigang Li
University of Florida

S089: Random matrix theory and applications

Room: Chi Wah Lecture Hall 206
Organizer: Shurong Zheng
Chair: Shurong Zheng

10:30 Tuning-Free Multiplier Bootstrap for Edge Eigenvalues of Large Covariance Matrices
◆Xiucai Ding, Yichen Hu, Jiahui Xie
University of California, Davis, University of California, Davis, University of California, Davis

10:55 Spectral Analysis of High-Dimensional Fisher Matrices Built from Sample Correlations
◆Xiaoyu Zhang, Shurong Zheng, Xue Ding
Northeast Normal University, Northeast Normal University, Jilin University

11:20 Extreme Eigenvalue Gaps in Random Matrices
Renjie Feng, Jiaming Li, ◆Dong Yao
The University of Sydney, Chinese Academy of Sciences, Jiangsu Normal university

S006: Advanced Statistical Methods for Network Data and Latent Variable Models

Room: Chi Wah Lecture Hall 115
Organizer: Haoran Zhang
Chair: Siliang Zhang

10:30 Network Goodness-of-Fit for the Block-Model Family
Jiashun Jin, Tracy Ke, Jiajun Tang, ◆Jingming Wang
Carnegie Mellon University, Harvard University, Harvard University, University of Virginia

10:55 Community Detection in Heterogeneous Signed Networks
Yuwen Wang, Shiwen Ye, ◆Jingnan Zhang, Junhui Wang
The Chinese University of Hong Kong, The Chinese University of Hong Kong, University of Science and Technology of China, The Chinese University of Hong Kong

11:20 Inference on Generalized Latent Variable Models with High-Dimensional Responses and Covariates
◆Jing Ouyang, Chengyu Cui, Yunxiao Chen, Kean Ming Tan, Gongjun Xu
University of Hong Kong, University of Michigan, London School of Economics and Political Science, University of Michigan, University of Michigan

11:55 Directed Network Discovery with Latent Variables: Identifiability and Estimation
◆Siliang Zhang, Yunxiao Chen, Irini Moustaki
East China Normal University, London School of Economics and Political Science, London School of Economics and Political Science

S094: Recent Advances in AI for Biological Sciences

Room: Chi Wah Lecture Hall 207
Organizer: Ruibin Xi
Chair: Ruibin Xi

10:30 Scaling Spatial Transcriptomics: Multiscale Virtual Modeling of Tissue Microenvironments
Zihao Chen
Peking University

10:55 Mapping the Landscape of Histomorphological Cancer Phenotypes with Self-Supervised Learning
Ke Yuan
University of Glasgow

11:20 Theory and Algorithms for Constructing AI-Based Dynamic Virtual Cells
Peijie Zhou
Peking University

S095: Recent Advances in Complex Data Analysis

Room: Chi Wah Lecture Hall 201

Organizer: Chen-Hung Kao

Chair: Chen-Hung Kao

10:30 Challenges and Solutions in Comparative Safety Profiles of COVID-19 Vaccines
Tsung-Shan Tsou
Graduate Institute of Statistics of Central University, Taiwan

10:55 Piecewise Fréchet Regression for Distributional Data
⬩*Jeng-Min Chiou, Yu-Ting Hsu*
Taiwan University & Academia Sinica, Taiwan,Taiwan University

11:20 Stacked Ensemble Prediction for Polynomial Regression with Heteroscedastic Measurement Errors
Jia-Ren Tsai
Fu Jen Catholic University

11:45 Integration of Individual Participant and Aggregate Data Under Dataset Shift
Ming-Yueh Huang
Academia Sinica, Taiwan

S101: Recent Advances in Longitudinal Data Analysis

Room: Chi Wah Lecture Hall 211

Organizer: Mengying You

Chair: Mengying You

10:30 Semiparametric Joint Modeling for Survival Analysis with Longitudinal Covariates
⬩*Wensheng Guo, Tianhao Wang*
University of Pennsylvania, Rush University

10:55 Varying Treatment Effects in Subgroups: A Unified Framework for Longitudinal Data Analysis
Pu Zhang, Xinsheng Zhang, Jiao Jin, ⬩*Jiakun Jiang*
Beijing Normal University, Beijing Normal University, Beijing Normal University, Beijing Normal University

11:20 Low-Rank Approximation and Detection for Structurally Changed Tensor SequenceSEQUENCE
Jiaqi Huang, ⬩*Ning Wang, Lixing Zhu*
Beijing Normal University

11:45 Low-Rank Spatio-Temporal State Space Models with Structured Spatial Covariance
⬩*Mengying You, Yuedong Wang, Wensheng Guo*
Shanghai University of International Business and Economics,University of California - Santa Barbara, University of Pennsylvania

S102: Recent Advances in Modeling Survival Data

Room: Chi Wah Lecture Hall 307

Organizer: Yingwei Peng

Chair: Yingwei Peng

10:30 Consistent Covariance Estimation for Stratum Imbalances Under Minimization Method for Covariate-Adaptive Randomization
Zixuan Zhao, Yanglei Song,⬩ *Wenyu Jiang, Dongsheng Tu*
Queen's University, Queen's University, Queen's University, Queen's University

10:55 Efficient Marginal Estimation for Clustered Survival Data with Informative Correlation Structure
Cheng Yang, Changcheng Li, Song Leng, Yingwei Peng, ⬩*Yi Niu*
Dalian University of Technology, Dalian University of Technology, Dalian Medical University, Queen's University, Dalian University of Technology

11:20 A Proportional Hazards Mixture Cure Model for Subgroup Analysis: Inferential Method and an Application to Colon Cancer Data
⬩*Kai Liu, Yingwei Peng, Narayanaswamy Balakrishnan*
Shanghai Lixin University of Accounting and Finance, Queen's University, McMaster University

11:45 Average Treatment Effect Estimation for Survival Data with a Cure Fraction
Hang Liu, ⬩*Yingwei Peng*
University of Electronic Science and Technology of China, Queen's University

S108: Recent Advances in Spatio-Temporal Data Analysis

Room: Chi Wah Lecture Hall 212

Organizer: Ching-Kang Ing

Chair: Shu-Hui Yu

10:30 On Lasso and Other Variable Selection Algorithms with Moderately Dense Coefficients
Ching-Kang ING, Yan-Shuo PAN, ⬩*CY (Chor-yiu) SIN*
Tsing Hua University, Taiwan, Tsing Hua University, Taiwan, Tsing Hua University, Taiwan

10:55 MDD-Based Estimation for Multivariate Spatial Temporal Models with Heterogeneous Coefficients
Muyi Li
Xiamen University

11:20 A Double Self-Normalized Approach to Change-Point Detection
Yaxing Yang
Xiamen University

11:45 Model Selection under Misspecification and Endogeneity: The MRIC Approach to Social Interaction Models
Hon-Ho Kwok
Taiwan University

S110: Recent Advances in Statistical Learning and Big Data

Room: Chi Wah Lecture Hall 301
Organizer: Xiaochao Xia
Chair: Xiaochao Xia

10:30 Variable Selection for Nadaraya-Watson-Frechet Regression
Yushen Dong, Youhan Lu, • Yichao Wu, Hans-Georg Muller
University of Illinois Chicago, University of Illinois Chicago, University of Illinois Chicago, University of California at Davis

10:55 Cluster Analysis of Longitudinal Associations Between Microbial Abundance and Environmental Factors
Emily Palmer, Caihong Qin, • Lan Xue, Yuan Jiang
HP Inc., Indiana University, Oregon State University, Oregon State University

11:20 Multiply-Penalized Decentralized Estimation with an Ap-plication to Convex Support Tensor Machines
Bei Zhang, Zemin Zheng, • Heng Lian
University of Science and Technology of China, City University of Hong Kong, University of Science and Technology of China, City University of Hong Kong Shenzhen Research Institute

S113: Recent Advances on Statistical Inference and Econometrics

Room: Chi Wah Lecture Hall 303
Organizer: Quefeng Li
Chair: Ruoqing Zhu

10:30 Factor-Adjusted Multiple Testing for High-Dimensional Individual Mediation Effects
Chen Shi, • Zhao Chen, Dan Wang
Fudan University, Fudan University, New York University Shanghai

10:55 Estimation of Rough Leverage Effect
Ming Cheng, • Christina Dan Wang,Ziyang Xiong
Fudan University, New York University Shanghai, Fudan University

11:20 A General Model Averaging Framework with Inference: Nested or Non-Nested Models
Yufei Li, • Fang Fang, Hansheng Wang
East China Normal University

11:50 Distributed Semi-Supervised Inference for Generalized Linear Models with Block-Wise Missing Covariates
Ziyuan Wang, Jin Liu, Jun Shao, Heng Lian, • Lei Wang
Nankai University, Nankai University, University of Wisconsin- Madison, City University of Hong Kong, Nankai University

S116: Recent Advances in Complex Time Series and Dynamic Network Analysis

Room: Chi Wah Lecture Hall 208
Organizer: Han Yan
Chair: Han Yan

10:33 Nonlinear Independent Component Analysis for Time Series
Han Yan
London School of Economics and Political Science

10:55 Change Point Detection and Identification in Tensor Factor Models
Zetai Cen
University of Bristol

11:20 Hypergraphs Embeddings: A Novel Approach with Increasing Dimensions
Xinyang Yu
London School of Economics and Political Science

11:45 Autoregressive Hypergraphs
Xianghe Zhu
London School of Economics and Political Science

S120: Robust and Flexible Statistical Methods for Complex and High-Dimensional Data

Room: Chi Wah Lecture Hall 304
Organizer: Xuejun Jiang
Chair: Hao Lei Weng

10:30 Robust Multi-Task Learning for Principal Component Analysis
Dali Liu, • Haolei Weng
Michigan State University, Southern University of Science and Technology

10:55 Double Penalized Variable Selection for High-Dimensional Partial Linear Mixed Effects Models
Yiping Yang, Chuanqin Luo, • Weiming Yang
Nanjing University Of Finance & Economics, Chongqing Technology and Business University, Chongqing Technology and Business University

11:20 Functional Additive Model via Convolution Smooth Quantile Estimation
• Ping Yu, Xiaojing Liu

Shanxi Normal University, Shanxi Normal University

11:45 Model-Free Rank Aggregation: A Maximum Score Approach
Haoran Zhang, Yunxiao Chen
Southern University of Science and Technology, London School of Economics and Political Science

S125: Statistical and Computational Advances for Single-Cell Genomics

Room: Chi Wah Lecture Hall 203
Organizer: Xinzhou Ge
Chair: Xinzhou Ge

10:30 High-Resolution Reconstruction of Single-Cell Specific Spatial Genome Architectures in 3D Space
Jianrong Wang
Michigan State University

10:55 CellScope: High-Performance Cell Atlas Workflow with Tree-Structured Representation
Bingjie Li, Runyu Lin, Tianhao Ni, Guanao Yan, Mannix Burns, Jingyi Li, Zhigang Yao
The Shanghai Institute for Mathematics and Interdisciplinary Sciences, National University of Singapore, Zhejiang University, Michigan State University, University of California, Los Angeles, Fred Hutchinson Cancer Center, National University of Singapore

11:20 SpaDOT: Optimal Transport Modeling Uncovers Spatial Domain Dynamics in Spatiotemporal Transcriptomics Studies
Wenjing Ma, Siyu Hou, Lulu Shang, Jiaying Lu, Xiang Zhou
The Chinese University of Hong Kong, Shenzhen, Yale University, The University of Texas MD Anderson Cancer Center, Emory University, Yale University

11:45 Controlling False Discoveries After Clustering via Data Splitting for Spatial Domain Marker Detection
Yingxin Lin, Lijun Wang, Hongyu Zhao
The Chinese University of Hong Kong, Zhejiang University, Yale University

S134: Statistical Inference for Complex Data Structures

Room: Chi Wah Lecture Hall 308
Organizer: Chiung-Yu Huang
Chair: Fan Xia

10:30 Estimating Causal Risk Ratio Under Complex Longitudinal Adherence Patterns with Time-Varying Principal Stratification
Rui Wang, Kwun Chuen Gary Chan, Fei Gao
University of Washington, University of Washington, Fred Hutchinson Cancer Center

10:55 Mediation with Random Effect Exposures
Fan Xia, Chuck McCulloch
University of California, San Francisco, University of California, San Francisco

11:20 Simultaneous Clustering and Estimation of Additive Shape Invariant Models for Recurrent Event Data
Zitong Zhang, Shizhe Chen
University of California, Davis, University of California, Davis

11:45 Reluctant Interaction Inference after Additive Modeling
Yiling Huang, Snigdha Panigrahi, Guo Yu, Jacob Bien
University of Michigan, University of Michigan, University of California Santa Barbara, University of Southern California

S152: Toward Efficient Clinical Trials: Statistical Methods That Leverage Baseline Covariates

Room: Chi Wah Lecture Hall 309
Organizer: Dong Xi
Chair: Zhiwei Zhang

10:30 To Adaptively Randomize or to Rerandomize: A Comparison of Covariate-Adaptive Randomization and Rerandomization
Ziji Qin, Feifang Hu, Yang Liu
The George Washington University, The George Washington University, Renmin University of China

10:55 Shape-Preserving Covariate Adjustment via Empirical Likelihood in Randomized Experiments
Zhilan Lou, Jun Shao, Tuo Wang, Yanyao Yi, Yu Du, Ting Ye
Zhejiang University of Finance and Economics, University of Wisconsin - Madison, Eli Lilly and Company, Eli Lilly and Company, Eli Lilly and Company, University of Washington

11:20 A Robust Score Test in G-Computation for Covariate Adjustment in Randomized Clinical Trials
Xin Zhang, Haitao Chu, Lin Liu, Satrajit Roychoudhury
Pfizer Inc., Pfizer Inc., Shanghai Jiao Tong University, Pfizer Inc.

11:45 Inverse Probability of Treatment Weighting: A Simple and Effective Approach to Covariate Adjustment for Survival Endpoints in Randomized Clinical Trials
Yongwu Shao, Zhiwei Zhang, Zhishen Ye
Gilead Sciences Inc., Gilead Sciences Inc., Gilead Sciences Inc.

S074: Modern Statistical Methods for Complex Longitudinal Data and Causal Inference

Room: Chi Wah Lecture Hall 310

Organizer: Xuejun Jiang

Chair: Xuejun Jiang

10:30 Longitudinal Mixed Membership Image-on-Scalar Model

Zhihao Wu, ◆ *Xinyuan Song, Kai Kang*

Yale University, The Chinese University of Hong Kong, Sun Yat-sen University

10:55 A Hierarchical Gaussian Process Framework for Longitudinal Ordinal Digital Survey Responses

Zoe Gibbs McBride, ◆ *Xiaojing Wang*

Brigham Young University,University of Connecticut

11:20 A Multiply Robust Estimation in Learning Optimal Individualized Treatment Regimes

Wensheng Zhu

School of Mathematics and Statistics, Yunnan University

11:45 A Unified Mendelian Randomization Framework for Inferring Causal Relationships: Accounting for Measured Covariates and Unmeasured Confounders

Xianghong Hu

School of Mathematical Science, Shenzhen University

S007: Advanced Statistical Modeling and Integration of Complex Data

Graphical Approaches for Multiplicity Adjustment in Clinical Trials

Jiangtao Gou
Villanova University

Graphical approaches provide highly flexible methods for multiplicity adjustment in clinical studies involving multiple endpoints. In particular, they make testing strategies transparent and easy to visualize. This talk presents several multiple testing procedures based on graphical methods.

Dynamic Representational Modeling of Language Learning and Processing

Gangyi Feng
The Chinese University of Hong Kong

We present a methodological framework for studying language learning and processing through dynamic representational modeling of neuroimaging data. The approach integrates MEG, fMRI, representational similarity analysis, cross-participant alignment, and task-trained language models to characterize how linguistic information is encoded and reorganized over learning. Across speech, speech-category learning, artificial-language acquisition, and generalization, we model neural activity in terms of evolving representational geometry and test its alignment with distributed cortical and corticostriatal population codes. This framework dissociates candidate learning computations, including prediction- and feedback-related signals, and quantifies their contributions to shared versus learner-specific neural structure. Speaker-listener analyses further assess whether language-specific representational organization is preserved across brains and predicts communicative success. By combining temporally resolved and spatially resolved neuroimaging with model-based analyses in a common representational space, this approach provides a compact and scalable method for tracking the emergence of linguistic knowledge across modalities, timescales, and individuals.

How Does Stress Shape Creativity? Evidence from Multimodal Data Analysis of Physiology, Cognition and Neuroscience

• *Haijun Duan, Yifan Wang*

Shaanxi Normal University, Shaanxi Normal University

Creativity is a uniquely human cognitive capacity and one of the most important human activities. At present, profound social changes pose serious challenges to human creativity. In the VUCA era, stress has become an inevitable condition of everyday life. Both individuals and organizations are frequently required to generate creative solutions under stress. Centered on the core question of whether stress undermines human creativity, the present research, drawing on a decade of systematic investigation, examines the mechanisms through which stress influences creativity, as well as potential intervention approaches, across physiological, psychological, neural, and behavioral levels.

At the individual level, the physiological basis of the effect of stress on creativity was first examined from the perspective of neuroendocrine changes. Salivary cortisol was used to assess activation of the hypothalamic–pituitary–adrenal (HPA) axis, while eye-blink rate and pupil diameter were used as indicators of dopaminergic and noradrenergic activity. The results showed that physiological responses under stress, including cortisol, eye-blink rate, and pupil diameter, negatively predicted creative performance, with cognitive flexibility serving as a mediator. In addition, EEG and fNIRS were employed to explore the cognitive-neural mechanisms underlying the effect of stress on creative activity. The findings indicated that stressed individuals showed sustained alpha desynchronization, lower right prefrontal activation, and weaker intra-prefrontal functional connectivity. At the group level, fNIRS hyperscanning was used to examine the effect of stress on group creativity from the perspective of intimate relationships and interpersonal interaction. The results showed that stress impaired group creativity. Groups characterized by the coexistence of heterogeneous personality traits and independent personality traits demonstrated better creative performance, showing a habitual response pattern of "tend-and-befriend." Romantic dyads exhibited significantly higher levels of cooperation and better creative performance than stranger dyads under stress, and males and females adopted different cooperative strategies in stressful situations. At the intervention level, post-stress transcranial direct current stimulation (tDCS) facilitated the alleviation of state anxiety, thereby improving stress-induced impairment in creativity. Pre-stress tDCS reduced negative emotional responses and anxiety changes under stress and alleviated creativity impairment by improving stress-related prefrontal dysfunction.

Furthermore, extending this line of research to individuals with drug addiction, the present study examined the cognitive–neural basis and intervention pathways of creative real-world problem solving, thereby advancing the understanding of the mechanisms underlying creative behavior. The results showed that individuals with drug addiction exhibited impaired creative real-world

problem solving, accompanied by reduced cognitive control, abnormal prefrontal activation, and impaired interpersonal cooperative interaction. Multimodal evidence from fNIRS, EEG, and eye-tracking further revealed abnormal brain function and altered neural interaction patterns in addicted individuals across both individual and collaborative contexts. Targeted and precise tDCS interventions based on impaired brain regions demonstrated both immediate and long-term improvements in creative real-world problem-solving ability in individuals with drug addiction, and further revealed the underlying neuromodulatory mechanisms. These findings converge with those observed in stress contexts across key domains, including cognitive control, prefrontal function, and cooperative interaction.

In summary, the present research integrates multimodal evidence to propose a comprehensive "biochemical–cognitive–neural–behavioral" framework and advances understanding of the mechanisms underlying both the impairment and enhancement of creativity, spanning stress contexts and special populations, and bridging mechanistic investigation with neuromodulatory intervention.

A Directed Weighted Graph Neural Network with Tensor Fusion for Multi-Omics Cancer Classification

Tiantian Yang
University of Idaho

Integration and analysis of multi-omics data provide valuable insights for improving cancer subtype classification. However, such data are inherently heterogeneous, high-dimensional, and exhibit complex intra- and inter-modality dependencies. Graph neural networks (GNNs) provide a principled framework for modeling these structures, but existing approaches often rely on prior knowledge or predefined similarity networks that produce either undirected or unweighted graphs, failing to capture task-specific directionality and interaction strengths. Interpretability at both the modality and feature levels also remains limited. To address these challenges, we propose TF-DWGNet, a novel Graph Neural Network framework that combines tree-based Directed Weighted graph construction with Tensor Fusion for multiclass cancer subtype classification. TF-DWGNet introduces two key methodological innovations: (i) a supervised tree-based strategy that constructs directed weighted graphs tailored to each omics modality, and (ii) a tensor fusion mechanism that captures unimodal, bimodal, and trimodal interactions using low-rank decomposition for computational efficiency. Experiments on three real-world cancer datasets demonstrate that TF-DWGNet consistently outperforms state-of-the-art baselines across multiple metrics and statistical tests. In addition, the model provides biologically meaningful insights through modality-level contribution scores and ranked feature importance. These results highlight that TF-DWGNet is an effective and interpretable solution for multi-omics integration in cancer research.

S010: Advances in High-Dimensional and Nonparametric Statistics

Conformal Prediction Beyond the Horizon: Distribution-Free Inference for Policy Evaluation

Feichen Gan, Youcun Lu, Yingying Zhang, • Yukun Liu
East China Normal University, East China Normal University, East China Normal University, East China Normal University

Reliable uncertainty quantification is crucial for reinforcement learning (RL) in high-stakes settings. We propose a unified conformal prediction framework for infinite-horizon policy evaluation that constructs distribution-free prediction intervals for returns in both on-policy and off-policy settings. Our method integrates distributional RL with conformal calibration, addressing challenges such as unobserved returns, temporal dependencies, and distributional shifts. We propose a modular pseudo-return construction based on truncated rollouts and a time-aware calibration strategy using experience replay and weighted subsampling. These innovations mitigate model bias and restore approximate exchangeability, enabling uncertainty quantification even under policy shifts. Our theoretical analysis provides coverage guarantees that account for model misspecification and importance weight estimation. Empirical results, including experiments in synthetic and benchmark environments like Mountain Car, show that our method significantly improves coverage and reliability over standard distributional RL baselines. This work was jointly completed by Feichen Gan, Youcun Lu, and Yingying Zhang.

Estimation and Inference for Density-Convoluted Support Vector Machine with Streaming Data

Haochen Rao, • Xu Guo, Heng Lian, Haobo Qi
Beijing Normal University, Beijing Normal University, City University of Hong Kong, Beijing Normal University

This paper investigates estimation and inference problems for the support vector machine coefficients with high-dimensional streaming data. To handle the non-smooth hinge loss, the density convolution technique is adopted. We first propose an online lasso estimator that can be obtained by optimizing a surrogate loss function. Instead of using complete historical data, the surrogate loss function uses a renewable quadratic form to approximate historical information. As a result, the estimation procedure only requires newly arrived data and limited historical information, which can be updated in an online manner. We derive the theoretical error bounds of the proposed online lasso estimator under mild conditions. To eliminate the inherent bias of the lasso estimator, we further propose an online debiased lasso estimator and construct a valid inference procedure. We establish the asymptotic normality of the debiased estimator. To numerically compute the proposed online lasso estimators, we consider the

proximal gradient descent algorithm, which is feasible for high-dimensional models and is computationally efficient. Extensive simulation studies and real-data analysis are conducted to illustrate the finite-sample performance of the proposed methods.

Rank Test of Independence Based on Adaptive Hellinger Distance

• *Wenwen Guo, Shilin Shang, Jiujing Wu*
Capital Normal University, Capital Normal University, Capital Normal University

We propose a novel independence measure for two random vectors, termed the Adaptive Hellinger Distance (AHD). This measure employs an innovative variant of the Hellinger distance (HD) and integrates methodological insights from the HHG test (introduced in Heller et al. (2012)). The proposed AHD test can be regarded as a crucial extension of the HHG test, as it shares a similar analytical framework while addressing certain limitations. Specifically, like the HHG test, it can detect dependence between two random vectors in any dimension, regardless of whether they are continuous or discrete. The rank test based on AHD performs exceptionally well and demonstrates robustness in applications. Furthermore, we have developed two tests utilizing p-value combination methods, named AHD-Simes and AHD-CCT. These methods are designed to integrate the results of tests conducted on four bins, ultimately enhancing overall performance. Simulation studies and real data analysis simultaneously illustrate the power and the utility of the proposed method, which is highly competitive especially when dealing with random vectors that are heavy-tailed or have outliers. We applied the proposed method to analyze two real datasets with different data types. Additionally, a new R package named AHD is developed to implement the proposed method.

Linear Spline Index Regression Model: Interpretability, Nonlinearity and Dimension Reduction

Lianqiang Qu, • *Long Lv, Meiling Hao, Liuquan Sun*
Central China Normal University, Central China Normal University, University of International Business and Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences

Inspired by the complexity of certain real-world datasets, this article introduces a novel flexible linear spline index regression model. The model posits piecewise linear effects of an index on the response, with continuous changes occurring at knots. Significantly, it possesses the interpretability of linear models, captures nonlinear effects similar to nonparametric models, and achieves dimension reduction like single-index regression models. In the model, both the locations and number of knots are unknown. This not only enhances the adaptability of the model in practical applications but also presents challenges in estimating and inferring the unknown parameters. Here, the knots represent points where the effects of covariates on the response exhibit significant changes while still maintaining continuity. Combining the penalized approach with convolution techniques, we propose a new method to simultaneously estimate the unknown parameters and the number of knots. The proposed method allows the number of knots to diverge with the sample size. We demonstrate that the proposed estimators can identify the number of knots with a probability approaching one and estimate the coefficients as efficiently as if the number of knots is known in advance. We also introduce a procedure to test the presence of knots. Simulation studies and two real datasets are employed to assess the finite sample performance of the proposed method.

S013: Advances in Modeling Frameworks for Structured and Complex Data

A Goodness-of-Fit Test for Zero-Inflated Poisson Mixed Effects Models

• *Juxin Liu, Yanyuan Ma, Jill Johnstone*
University of Saskatchewan, Pennsylvania State University, University of Saskatchewan

We propose a new goodness-of-fit (GOF) test motivated by a study of tree seedling abundance after forest fires. The data exhibits common challenges in ecological research, including hierarchical count structures, excess zeros, and unbalanced repeated measures. Although Zero-Inflated Poisson (ZIP) mixed-effects models are widely used to handle these challenges, a robust omnibus GOF test has been lacking. We develop a test statistic based on cumulative sums of residuals and establish its theoretical properties, including consistency and asymptotic convergence. The application of the proposed test is illustrated by a real data example and simulation studies.

Combined Estimator Using Random Forests and Support Vector Machine in Regression

• *Yan Lu, Ning Xie*

University of New Mexico, Michigan State University
When the relationship between the response and predictors is nonlinear,machine learning tools such as random forests and support vector regression are highly effective for regression analysis. To leverage the strengths of both estimators,we propose a novel combined weighted estimator that optimizes prediction accuracy through the careful selection of weights. Our proposed combined estimator demonstrates superior performance compared to existing methods for regression analysis, as supported by simulation results and hypothesis tests.

A Semi-Parametric Model for Small Area Estimation Using

Support Vector Machine

• *Guoyi Zhang, John Pleis*
University of New Mexico, National Center for Health Statistics

The Fay-Herriot model (Fay & Herriot, 1979) is popularly used in small area estimation (SAE). Fay-Herriot model assumes a linear relationship between the response and predictor variables, which in practice may not be appropriate. In this research, we extend the Fay-Herriot model to a more general semi-parametric model for small area estimation (SP-SAE), in which the nonparametric component is estimated by machine learning tools such as support vector machine. A back-fitting algorithm is proposed to solve the SP-SAE problem. An example using the American Community Survey (ACS) data shows that the proposed SP-SAE significantly improved the health insurance coverage estimators than the Fay-Herriot model.

S023: Advances in Time Series Modeling and Functional Data Analysis

Testing and Estimation of Change Point in ARMA Model With Heavy-Tailed G-GARCH Noises

• *Qiang Bai, Shiqing Ling, Ke Zhu*

Shanxi University of Finance and Economics, Hong Kong University of Science and Technology, The University of Hong Kong

We develop a procedure for testing and estimating the change point in the autoregressive moving average (ARMA) model with heavy-tailed general generalized autoregressive conditional heteroskedasticity (G-GARCH) noises. Based on the self-weighted least absolute deviation estimator (SLADE), we propose two score-type test statistics for change point detection. Under the null hypothesis, we show that one of them converges weakly to the maxima of a Brownian bridge, and the other one converges weakly to an extreme distribution. Furthermore, we prove that the SLADE of the change point converges weakly to the location of the maxima of a double-sided random walk, and the SLADE of other parameters is asymptotically normal. Our two score-type tests and SLADE are applicable for the data with an infinite variance, while not specifying the form of G-GARCH noises. Finally, we use simulations and two real examples to demonstrate the usefulness of the proposed tests and estimator in handling the heavy-tailed data with an infinite variance.

A Class of Adaptive MIDAS Linear Regression Models

• *Xingfa Zhang, Mengli Liu, Keke Hu*
Guangzhou University, Guangzhou University, Guangzhou University

This paper investigates regression models incorporating data sampled at mixed frequencies. We establish a theoretical lower bound for the asymptotic variance of mixed-frequency regression estimators and develop a novel data-driven weighting scheme that adaptively optimizes frequency-specific weights, overcoming the constraints inherent in conventional fixed or parametric weighting frameworks. The proposed methodology is rigorously examined through theoretical analysis, comprehensive Monte Carlo simulations, and empirical applications, demonstrating its statistical efficiency and practical utility in mixed-frequency regression contexts.

Variable Selection and Estimation for Generalized Linear Models with Longitudinal Predictors

• *Xiaojuan Han, Lizhe Sun, Xiaoqin Zhang, Xiang Zhan*
Shanxi University of Finance and Economics, Shanxi University of Finance and Economics, Shanxi University of Finance and Economics, Southeast University

Generalized varying-coefficient models for longitudinal data play a crucial role in modern statistical research, with broad applications in both social and medical sciences. While most existing methods assume that both the predictors and the response are longitudinal, this presentation introduces a novel framework of generalized varying-coefficient models tailored to settings where the covariates are longitudinal but the response is static. We employ B-spline techniques to approximate the varying-coefficient functions and propose variable selection and estimation procedures based on a group L0 penalty within this framework. Simulation results demonstrate that the proposed method consistently outperforms existing approaches in terms of both prediction accuracy and estimation efficiency. Finally, to investigate risk factors for diabetes, we apply the proposed method to data from the China Health and Retirement Longitudinal Study (CHARLS).

Wavelet-Factor Matrix Hybrid Transport Maps for Spatio-temporal Bayesian Fusion

Shiyuan He
Beijing Technology and Business University

Acquiring remote sensing images with both high spatial resolution and high temporal frequency is crucial for monitoring dynamic land-surface processes. However, classical Bayesian methods, despite their rigor, are constrained by the curse of dimensionality and slow convergence of iterative procedures.To overcome these challenges, we propose a Wavelet-Factor Matrix Hybrid Transport Maps for Spatiotemporal Bayesian Fusion(WaFT). WaFT decomposes images via two-dimensional discrete wavelet transform (2D-DWT) into Stationary high-frequency Spatial Details (SSD) and Dynamic low-frequency Large-sacle Structures (DLS). For the dynamic components, we introduce a low-rank matrix factorization model that projects high-dimensional

dynamics into a compact latent subspace, effectively capturing phenology evolution. Unlike iterative optimize estimation or deep learning training, WaFT derives closed-form analytical solutions directly through conjugate Gaussian formulas, enabling efficient, training-free inference. Comprehensive experiments on agricultural (CIA) and urban (Tianjin) datasets demonstrate that WaFT outperforms the compared methods in RMSE, SSIM, and SAM metrics, preserving spectral fidelity while maintaining spatial structural integrity, and showing robustness against sensor noise variations.

S028: Bayesian Regression and Variable Selection for Structured and High-Dimensional Data

Integrative Bayesian Modeling for Functional MRI Data

Xuan Cao
University of Cincinnati

In this talk, I will present several statistical methods for analyzing functional MRI data aimed at improving disease prediction and diagnosis. Joint estimation of multiple brain networks is used to identify aberrant functional connectivity associated with neurological disorders, while network-guided feature selection reveals abnormal connectivity patterns and radiomics-based functional insights. In addition, a sequential Bayesian approach is introduced for estimating effective and functional connectivity using vector autoregressive models. This framework incorporates pseudo-likelihood methods and a novel refitting procedure to enable efficient inference in high-dimensional settings. The proposed methods are supported by theoretical guarantees and demonstrate improved sensitivity, precision, and interpretability. Together, these approaches provide new insights into connectivity structures and potential biomarkers for neurodegenerative diseases.

Bayesian Region Selection and Prediction in Poisson Regression with Spatially Dependent Global-Local Shrinkage Prior

Zihan Zhu, ⋆Xueying Tang, Shuang Zhou
Yale University, University of Arizona, Arizona State University

High-dimensional spatially correlated covariates are common in regression models encountered in environmental sciences and other fields. In such models, the regression coefficients often exhibit a sparse structure with spatial dependence. Although standard variable selection approaches can help detect the sparse structure, incorporating the dependence into variable selection helps recover spatially contiguous signals and improves prediction accuracy. Motivated by a real-world challenge in hurricane count prediction, we propose a novel neighborhood-structured global-local shrinkage prior for prediction and region selection in Poisson regression with spatial covariates. The proposed prior combines the Conditional Auto-Regressive (CAR) prior with a Super Heavy-tailed prior to introduce spatial dependence among the coefficients while ensuring appropriate shrinkage effects for covariate selection. We develop an efficient Metropolis-within-Gibbs sampler for computation that accommodates the count data. Extensive simulation studies demonstrate that the proposed model excels when signals are weak and adjacent and the spatial dependence in covariates is strong. In the application of hurricane prediction from the north Atlantic, our method outperforms traditional regression-based approaches and rivals the benchmark "oracle" model.

BayesBadgeR: Bayesian Double Generalized Beta Regression with Spatial and Temporal Borrowing for Bounded Well-Being Indices

Abhi Jain,Michael LaValley,Kimberly A. Dukes, ⋆ Shariq Mohammed
Boston University, Boston University, Boston University,Boston University

Bounded outcomes arising from health and well-being indices present unique modeling challenges, particularly when interest lies in both the mean and dispersion structures and in inference at fine spatial scales. In this talk, I introduce BayesBadgeR, a Bayesian double generalized beta regression framework for analyzing a bounded individual-level well-being index (WBI) using repeated cross-sectional survey data collected in Massachusetts from 2021 to 2023. The proposed model jointly specifies regression structures for both the mean and precision parameters of the beta distribution. To account for spatial dependence, we incorporate a graph Laplacian penalty induced by a driving time–based neighborhood structure across ZIP Codes. Temporal dependence is introduced through a sequential Bayesian updating strategy, whereby posterior distributions of spatial effects from one year inform priors for subsequent years. This induces a principled form of temporal borrowing despite the absence of longitudinal follow-up at the individual level. This unified spatial–temporal borrowing framework improves parameter estimation and stabilizes inference in sparsely sampled regions. Through a realistic simulation study, we demonstrate gains in estimation efficiency and reduced uncertainty relative to models that ignore dependence structures. We then apply the method to Massachusetts WBI data, identifying income, educational attainment, and marital status as key predictors of individual-level well-being, while spatial effects reveal systematic geographic variation, with higher well-being concentrated in Western Massachusetts, Cape Cod, and areas surrounding Boston. Overall, BayesBadgeR provides a flexible and interpretable framework for modeling bounded outcomes with complex dependence structures, with broad applicability to small-area estimation problems in public health and beyond.

S031: Bridging Statistics and Modern AI: Foundations for Deep Learning and Generative Models

Statistical Inference in an Interactive Learning Paradigm

⬩ *Yuchen Wu, Kangjie Zhou, Weijie Su*
Cornell University, Columbia University, University of Pennsylvania

The proliferation of generative artificial intelligence has given rise to an interactive learning paradigm, where model parameters are continuously updated using not only data generated by natural processes, but also synthetic outputs produced by other models. This paradigm introduces two major challenges: (1) training data are no longer drawn exclusively from the target population, undermining a core assumption of classical statistical learning theory, and (2) model training processes become inherently correlated, as models influence one another through repeated exposure to each other's synthetic outputs. Establishing reliable statistical inference in such interactive environments therefore remains an important open problem. In particular, there is growing concern about model collapse, a phenomenon in which the performance of generative models progressively degrades as they are trained on synthetic data produced by earlier model generations.

In this work, we study the behavior of generative models under general interaction patterns. We formalize these interactions using directed graphs and show that the occurrence of model collapse depends critically on the graph's topology. Within this framework, we derive an explicit necessary and sufficient condition characterizing when model collapse occurs. Our analysis covers both finite-sample results for linear regression and asymptotic guarantees for general M-estimators. We further validate our theoretical findings through extensive numerical experiments.

Transformers Meet In-Context Learning: A Universal Approximation Theory

Gen Li, Yuchen Jiao, ⬩ *Yuting Wei, Yuxin Chen*
Chinese University of Hong Kong, Chinese University of Hong Kong, University of Pennsylvania, University of Pennsylvania

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.

Generative Adversarial Learning with Optimal Input Dimension and Its Adaptive Generator Architecture

Zhiyao Tan, Ling Zhou, ⬩ *Huazhen Lin*
Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, Southwestern University of Finance and Economics

In this talk, we investigate the impact of the input dimension on the generalization error in generative adversarial networks (GANs). We first provide both theoretical and practical evidence to validate the existence of an optimal input dimension (OID) that minimizes the generalization error. Then, to identify the OID, we introduce a novel framework called generalized GANs (G-GANs), which includes existing GANs as a special case. By incorporating the group penalty and the architecture penalty developed in the paper, the proposed G-GANs have several intriguing features. First, our framework offers adaptive dimensionality reduction from the initial dimension to a dimension necessary for generating the target distribution. Second, this reduction in dimensionality also shrinks the required size of the generator network architecture, which is automatically identified by the proposed architecture penalty. Both reductions in dimensionality and the generator network significantly improve the stability and the accuracy of the estimation and prediction. Theoretical support for the consistent selection of the input dimension and the generator network is provided. Third, the proposed algorithm involves an end-to-end training process, and the algorithm allows for dynamic adjustments between the input dimension and the generator network during training, further enhancing the overall performance of the G-GANs. Extensive experiments conducted with simulated and benchmark data demonstrate the superior performance of the proposed G-GANs. Moreover, the features generated based on the input dimensions identified by G-GANs align with visually significant features.

Deep Discrete Encoders: Identifiable Deep Generative Models for Rich Data with Discrete Latent Layers

Seunghyun Lee, ⬩ *Yuqi Gu*
Columbia University, Columbia University

In the era of generative AI, deep generative models (DGMs) with latent representations have gained tremendous popularity.

Despite their impressive empirical performance, the statistical properties of these models remain underexplored. DGMs are often overparametrized, non-identifiable, and uninterpretable black boxes, raising serious concerns when deploying them in high-stakes applications. Motivated by this, we propose interpretable deep generative models for rich data types with discrete latent layers, called Deep Discrete Encoders (DDEs). A DDE is a directed graphical model with multiple binary latent layers. Theoretically, we propose transparent identifiability conditions for DDEs, which imply progressively smaller sizes of the latent layers as they go deeper. Identifiability ensures consistent parameter estimation and inspires an interpretable design of the deep architecture. Computationally, we propose a scalable estimation pipeline of a layerwise nonlinear spectral initialization followed by a penalized stochastic approximation EM algorithm. This procedure can efficiently estimate models with exponentially many latent components. Extensive simulation studies for high-dimensional data and deep architectures validate our theoretical results and demonstrate the excellent performance of our algorithms. We apply DDEs to three diverse real datasets with different data types to perform hierarchical topic modeling, image representation learning, and response time modeling in educational testing.

S033: Computational and Statistical Advances in Association Modeling

Adjusting for Non-Confounding Covariates in Case-Control Association Studies

Siliang Zhang, Jinbo Chen, Zhiliang Ying, ⋆ *Hong Zhang*
East China Normal University, University of Pennsylvania, Columbia University, University of Science and Technology of China

There is a substantial literature in case-control logistic regression on whether or not non-confounding covariates should be adjusted for. However, only limited and ad hoc theoretical results are available on this important topic. A constrained maximum likelihood method was recently proposed, which appears to be generally more powerful than logistic regression methods with or without adjusting for non-confounding covariates. This note provides a theoretical clarification for the case-control logistic regression with and without covariate adjustment and the constrained maximum likelihood method on their relative performances in terms of asymptotic relative efficiencies. We show that the benefit of covariate adjustment in the case-control logistic regression depends on the disease prevalence. We also show that the constrained maximum likelihood estimator gives an asymptotically uniformly most powerful test.

Integrating Network Annotation From Multiple Correlated Traits To Improve Polygenic Risk Scores Based on GWAS Summary Statistics

Lirong Zhu, Xuewei Cao, Shuanglin Zhang, ⋆ *Qiuying Sha*
Tianjin Medical University, Columbia University, Michigan Technological University, Michigan Technological University

Polygenic risk scores (PRS) are valuable tools for predicting disease risk based on genetic information, with potential impacts on disease prevention and early treatment strategies. Although thousands of disease-associated genetic variants have been identified through genome-wide association studies (GWAS), the accuracy of genetic risk prediction for most diseases remains moderate and challenging. In this paper, we introduce NetPRS, a novel method that utilizes a penalized regression model and leverages network annotation information to enhance PRS prediction. This network annotation is obtained from a genotype-phenotype bipartite network (GPN), where multiple SNPs and traits are linked based on association strengths obtained from GWAS summary statistics. The network annotation allows for the incorporation of information from relevant traits into the PRS prediction for the target trait. Compared to state-of-the-art risk prediction methods, NetPRS consistently achieves improved prediction accuracy in both simulation studies and real data analysis.

Sparse CCA-Based Mediation Analysis with High-Dimensional Exposures and Mediators

Xincheng Li, ⋆ *Hongmei Jiang*
Northwestern University, Northwestern Univeresity

Mediation analysis is essential for elucidating how exposure variables affect health outcomes through intermediate variables, or mediators, in environmental studies. In settings involving numerous environmental exposures (e.g., chemical mixtures or pollutants) alongside multiple potential mediators (e.g., metabolites), advanced methods are required to effectively disentangle direct and indirect effects. In this paper, we introduce a novel mediation analysis approach based on Sparse Canonical Correlation Analysis (SCCA), tailored for high-dimensional exposures and mediators. The performance of the proposed method is demonstrated through both simulation studies and an application to real-world data.

Cluster Analysis of Longitudinal Associations Between Microbial Abundances and Environmental Factors

Emily Palmer, Caihong Qin, Lan Xue, ⋆ *Yuan Jiang*
Hewlett-Packard, Indiana University, Oregon State University, Oregon State University

Microbial communities are dynamic systems that shift in composition in response to host and environmental factors.

Understanding these temporal patterns is essential for identifying taxa with shared biological roles. However, standard clustering approaches for microbiome data often neglect its longitudinal structure and fail to account for external variables that may influence microbial trajectories. We introduce a novel clustering framework named \textit{cleverly}, denoting (Cl)ustering with (E)xternal (V)ariabl(e)s in (R) with (L)ongitudinal (Y)s, which incorporates external variables into the clustering of multivariate responses represented as longitudinal and compositional counts. Our method enables the generation of clusters based either on the responses' baseline trajectories or on their reactions to external perturbations, as well as accommodates the compositional, discrete, and time-correlated nature of the data. Using simulated data, we demonstrate that \textit{cleverly} recovers true groupings under a range of biologically realistic conditions, including varying correlation structures and compositional constraints. We further apply \textit{cleverly} to longitudinal fecal microbiome data from wild African buffalo (\textit{Syncerus caffer}), revealing distinct clusters of taxa that either maintain consistent trajectories or respond sharply to changes in resource availability. These findings underscore the importance of accounting for environmental covariates in microbiome clustering and illustrate how \textit{cleverly} enables a deeper understanding of microbial function and host-microbiome interactions in dynamic ecosystems.

S036: Data Integration Under Different Data Types: New Methodology Developments

A Semiparametric Two-Sample Homogeneity Test with Nonignorable Nonresponse Using Callback Data

Xinyu Wang, Tao Yu, ◆Chunlin Wang, Pengfei Li

Xiamen University, National University of Singapore, Xiamen University, University of Waterloo

Testing the homogeneity of two distributions is fundamental in statistics, but classical procedures may fail under nonignorable nonresponse. In many surveys, callback data record repeated contact attempts and provide auxiliary information about the response mechanism. We develop a semiparametric framework for two-sample homogeneity testing that explicitly incorporates such information. The response mechanism is modeled by a flexible semiparametric callback model, while the two population distributions are linked through a density ratio model. Within this unified framework, we propose an empirical likelihood ratio test for distributional homogeneity and show that, under the null hypothesis, it has a Wilks-type chi-square limit. To facilitate computation, we develop an efficient expectation-maximization-type algorithm. Simulation results show that the proposed method controls type I error well and achieves substantially higher power than existing methods that ignore nonignorable missingness. An application to real survey income data illustrates its practical value.

Orthogonal Covariate Balancing for Causal Inference

Ying Yan

Sun Yat-sen University

In this work, we propose a maximum mean discrepancy framework for balancing covariate distributions in different treatment groups. By the theory of reproducing kernel Hilbert spaces, we show that important covariate moments are guaranteed to be exactly balanced, and meanwhile all other aspects of covariate distributions can be approximately balanced. Moreover, we reveal that the proposed framework is closely related to kernel ridge regression by the duality theory in optimization. Low rank approximation is used to reduce computational cost. Theory and numerical studies will be presented.

Bias-Corrected Byzantine-Robust Estimator via Cornish-Fisher Expansion for Distributed Learning

Zhonglei Wang

Xiamen University

For distributed learning, median-of-means estimators are popular for statistical inference against Byzantine attacks, but they suffer from inefficiency or even bias, when the sample is asymmetrically distributed. We tackle this problem by proposing a bias-corrected Byzantine-robust estimator via Cornish-Fisher expansion, and it can be widely implemented in various gradient descent algorithms for state-of-the-art deep learning models. We demonstrate that the bias of the proposed estimator is much smaller than a standard median-of-means estimator and its variations under regularity conditions, and asymptotic properties of the proposed estimator are also established. The proposed estimator outperforms its alternatives numerically in terms of mean absolute error under different synthetic setups, and performs the best on both regression and classification tasks when analyzing the LendingClub loan dataset and the Stroke Prediction dataset, respectively.

Specification Testing in Complex Survey Sampling With Missing Data

◆*Puying Zhao, Mengmeng Xu*

Yunnan University, Yunnan University

This study develops a unified framework for specification testing in linear quantile models with complex survey designs and missing data. Under the assumption that the missingness mechanism is missing at random, we propose a nonparametric orthogonal projection-based test to examine the correct specification of a linear conditional quantile function over

a continuum of quantile levels. The proposed framework enables valid post-design-based inference on the effects of covariates across the conditional distribution of outcomes, while simultaneously accounting for complex survey design features and alleviating biases induced by missing observations. We derive the asymptotic distribution of the test statistic under both the null and alternative hypotheses and show that parameter estimation has no asymptotic effect on its limiting distribution. For practical implementation, a multiplier bootstrap procedure is developed and its theoretical validity is established. The finite-sample performance of the proposed method is investigated through simulation studies, and its practical utility is illustrated with a real-world application.

S038: Emerging Statistical Frameworks for High-Dimensional and Biased Data

Simulation-Based Inference via Structured Score Matching

Haoyu Jiang, ◆ Yuexi Wang, Yun Yang

University of Illinois Urbana-Champaign, University of Illinois Urbana-Champaign, University of Maryland, College Park

Simulation-based inference (SBI) provides a powerful framework for statistical analysis when the likelihood function is intractable but simulations are available. I will present a unified framework that leverages structured score matching to enable both maximum likelihood estimation and Bayesian posterior sampling in such likelihood-free settings. The key idea is to approximate the intractable likelihood score function with neural networks that embed the statistical structures of likelihood score through architectural regularization. These structural constraints enhance estimation accuracy, scalability, and uncertainty quantification. We establish theoretical guarantees for the proposed methods and demonstrate their practical advantages on benchmark tasks and challenging high-dimensional problems, where they perform favorably against existing approaches.

High-Dimensional Error-in-Variables Survival Analysis

Jiarui Zhang, Hongsheng Liu, Xin Chen, ◆ Jinfeng Xu

South China of Technology, The University of North Carolina at Chapel Hill, South China University of Science and Technology, City University of Hong Kong

High-dimensional survival data with measurement errors are prevalent in biomedical studies, where numerous clinical or genetic variables are collected for risk assessment. The presence of measurement errors in covariates substantially complicates parameter estimation and variable selection, often leading to challenging non-convex optimization problems. We propose an error-in-variables additive hazards regression model for high-dimensional noisy survival data. By employing the nearest positive semi-definite matrix projection, we develop a fast Lasso approach and its soft-thresholding variant, both supported by strong theoretical guarantees. Under mild assumptions, we establish model selection consistency, oracle inequalities, and asymptotic distributions for the proposed methods. Simulation studies and applications to two real datasets demonstrate the superior performance of these methods in handling high-dimensional data. Notably, they exhibit remarkable robustness even in scenarios with missing values, highlighting their practical utility in complex biomedical settings.

Addressing Confounding in Comparative Effectiveness Research Using Observational Multilevel Semi-Competing Risks Data

Hyunjae Cho, ◆ Hong Zhu

University of Virginia, University of Virginia

Population-based comparative effectiveness research (CER) using observational healthcare data, is a primary research tool for assessing treatment effectiveness. However, the complex nature of these data necessitates the development of novel methodology to obtain valid results. Particularly, a ‘semi-competing risks’ situation arises when the observation of non-terminal (e.g., cancer recurrence) is subject to a terminal event (e.g., death), but not vice versa. The analysis of such data is further complicated by the multilevel data structure and confounding bias. To address these complexities, we propose a copula-based, hierarchical semi-competing risks regression modeling framework and incorporate propensity score (PS) weighting to estimate average treatment effects (ATEs) on non-terminal and terminal events. We propose a two-stage estimation procedure and establish its asymptotic properties. The proposed method is evaluated through extensive simulation studies and applied to a multi-center studies of breast cancer.

Capturing High-Dimensional Brain Atrophy Patterns via MRI Distance Geometry and Cross-Attention

Zheyu Wang

Johns Hopkins University

Early detection of Alzheimer’s Disease (AD) during the preclinical stage is critical, as brain changes are minimal and therefore offer a vital therapeutic window for preserving cognitive function. Cerebrospinal fluid (CSF) biomarkers show high diagnostic accuracy in this phase but the need for lumbar puncture limit their utility for widespread screening. MRI, on the other hand, is non-invasive, widely accessible and increasingly used in AD diagnosis. However, its performance in distinguishing Mild Cognitive Impairment (MCI) from cognitively normal individuals in the early phases is suboptimal. While incorporating known risk factors such as age and ApoE4 status improves diagnostic performance, the added value of MRI in early detection remains

limited. This leads to a crucial question: is this limitation inherent to MRI's temporal placement within the biomarker cascade (e.g., typically changing after amyloid pathology but before major cognitive decline), or because its full diagnostic potential is currently underutilized?

In this work, we explore two novel approaches to enhance MRI's capability for early AD detection. First, we recognize that atrophy may have different rate of change across brain region, and the traditional use of MRI volumetric measures may not fully capture the multidimensional nature of degeneration. We propose high-dimensional distance metrics to better represent complex patterns of brain atrophy over time, and use a two-step "wisdom-of-crowds" ensemble strategy to aggregate information across regions and metrics for improved classification. Second, we investigate a deep learning approach that directly analyzes 3D MRI scans. To improve performance with limited sample sizes, we utilize a cross-attention mechanism that integrates Jacobian determinant maps (JSM) with intensity maps. This fusion enables the model to leverage spatial deformation patterns and detailed anatomical structures, enhancing the detection of subtle brain changes associated with AD. Results from both approaches demonstrate enhanced performance compared to traditional methods, showing promise for MRI as an early detection modality in the early phases of AD.

S039: Estimation and Testing in Sparse, Quantile, and Survival Models

Orthogonalized Kernel Debiased Machine Learning for Logistic Multimodal Inference

Yaohua Rong

Beijing University of Technology

Multimodal imaging has reshaped Oncology research. Notwithstanding the unprecedented opportunities it affords, this development is nevertheless counterbalanced by considerable challenges. In particular, within the framework of logistic multimodal regression models, it is difficult to combine the dual merits of the interpretability inherent to simple association models and the flexibility afforded by highly adaptive nonlinear models. In this article, we propose logistic orthogonalized kernel debiased machine learning method, which employs Neyman orthogonality to construct nonlinear score functions, while adopting a decomposition-based orthogonality strategy to resolve model identifiability issues. We target the prevalent multimodal setting comprising a primary modality of interest alongside auxiliary modalities. It is established that the estimated primary parameter is both root-N-consistent and asymptotically normal, and is, in addition, semi-parametrically efficient. Under the logistic regression framework, our proposed method distinguishes itself from existing statistical approaches for multimodal data fusion and orthogonalized machine learning methods for high-dimensional inference, by achieving model interpretability and model flexibility simultaneously. We demonstrate the efficacy of our method through extensive simulations and an application for lung cancer detection using radiomics data derived from multimodal medical imaging. Supplementary materials for this article are accessible online.

Distributed Hard Screening for Quantile Regression

Xianhui Liu

Jiangxi University of Finance and Economics

Nowadays, mass data usually is distributionally stored in multiple datasets. For high dimension data, feature screening is a benchmark procedure. In this paper, a distributed hard-screening is proposed for quantile regression-a robust alternative to least square regression. The proposed method is implemented via Alternating Direction Method of Multipliers, and the algorithmic convergence is established. Furthermore, the method is shown to have sure-screening property, under some mild conditions. The performance of the method is illustrated with numerical simulations and a real data application.

SID: A Novel Class of Nonparametric Tests of Independence for Censored Outcomes

Jinhong Li, Jicai Liu, Jinhong You, Riquan Zhang

Zhejiang Gongshang University, Shanghai Lixin University of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai University of International Business and Economics

We propose a new class of metrics, called the survival independence divergence (SID), to test dependence between a right-censored outcome and covariates. A key technique for deriving the SIDs is to use a counting process strategy, which equivalently transforms the intractable independence test due to the presence of censoring into a test problem for complete observations. The SIDs are equal to zero if and only if the right-censored response and covariates are independent, and they are capable of detecting various types of nonlinear dependence. We propose empirical estimates of the SIDs and establish their asymptotic properties. We further develop a wild bootstrap method to estimate the critical values and show the consistency of the bootstrap tests. The numerical studies demonstrate that our SID-based tests are highly competitive with existing methods in a wide range of settings.

S042: From Cells to Tissues: Statistical Frameworks for Spatial Omics

Digital Tissue Twins: Uniting Spatial Omics with Histopathology

Mingyao Li

University of Pennsylvania

Spatial omics technologies are reshaping how we study the molecular organization of tissues, offering unprecedented opportunities to map biology in its native spatial context.However, key challenges remain in resolution, scalability, and accessibility. In this talk, I will describe how my lab is developing AI-powered approaches that integrate spatial omics with histopathology to overcome these barriers. I will highlight computational strategies that make spatial omics experiments smarter, faster, and more informative, from enhancing spatial resolution beyond current experimental limits, to extending analyses across large and complex tissues, to integrating diverse molecular measurements within a unified framework. Together, these approaches enable the creation of "digital tissue twins,"computationalm representations of tissues that capture both molecular and structural organization. Finally, I will discuss our efforts to design AI systems that are transparent and trustworthy, ensuring that these tools can support robust biological discovery and clinical translation. Theseadvances aim to unlock the full potential of spatial omics and enable new insights in the era of precision medicine.

Decoupling Biological Signal from Technical Confounding in Library Size for Single Cell and Spatial Omics

Kedi Le, Weiqiang Zhou, ⋆Hongkai Ji

Johns Hopkins Bloomberg School of Public Health, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins Bloomberg School of Public Health

Library size is routinely treated as a technical confounder in single-cell and spatial omics workflows, with normalization applied early to remove presumed technical variation prior to downstream analysis. Recent work has shown, however, that library size in spatial transcriptomics can also reflect underlying tissue structure, and that standard normalization can obscure biologically meaningful signals and degrade tasks such as spatial domain detection. Here, we extend this observation to additional modalities, demonstrating that library size carries biologically relevant information in spatial ATAC-seq and single-cell RNA-seq as well. We show that naive library size normalization can mask important biology and impair downstream interpretation. We evaluate various normalization strategies and present a solution to separate biologically informative variation from technical confounding in library size, preserving meaningful signal while mitigating unwanted technical effects.

A Unified Framework for Extracting Single-Cell Information from High-Resolution Spatial Transcriptomics

Can Yang

Hong Kong University of Science and Technology

Understanding how different cell types assemble into tissues and organs, as well as how they interact to transmit and receive biological signals, is essential for advancing biomedical and biological research. Recent advancements in spatial transcriptomics (ST) technologies have opened new avenues for investigating biological systems by achieving subcellular spatial resolution. Since cells are the fundamental units of life, extracting single-cell information from high-resolution ST data is crucial. However, existing ST platforms often capture sparse transcript counts per spot or measure only a limited number of genes, complicating the extraction of comprehensive single-cell information. In this study, we introduce CellART, a unified framework designed to extract single-cell information across diverse high-resolution ST platforms, including VisiumHD, Xenium, MERFISH, and Stereo-seq. By leveraging multimodal data, such as staining images, spatial transcriptomics data, and single-cell RNA sequencing references, CellART simultaneously performs cell segmentation and cell type annotation through a seamless integration of deep learning and probabilistic modeling. We demonstrate the efficiency, generalizability, and robustness of CellART across various high-resolution spatial transcriptomics platforms, capable of processing datasets containing millions of spots. Comprehensive experiments validate the biological relevance and accuracy of the recovered cellular information within spatial configurations. Notably, we highlight the utility of CellART in breast and colorectal cancer datasets, showcasing its ability to fully leverage high-resolution ST data. By enhancing cellular resolution, CellART facilitates the identification of transient cancer cell states and immune cell subtypes. Furthermore, CellART enables investigations into cancer-immune cell communication, uncovering both established interactions and novel ligand-receptor pairs. The outputs of CellART are compatible with widely used community tools, facilitating a variety of downstream analyses.

CoxFormer Enables Spatial Omics Inference with Multimodal Generative Modeling

⋆*Jin Liu, Yiyang Yang*

The Chinese University of Hong Kong□Shenzhen, The Chinese University of Hong Kong□Shenzhen

Gene co-expression provides a principled basis for transcriptome-wide gene–gene relationships, yet the available high-quality estimates typically cover less than half of the genome. In the current biotechnological landscape, spatial omics faces complementary measurement constraints: targeted \textit{in situ} platforms restrict gene coverage, whereas sequencing-based assays lack cellular resolution. When extended to the

entire transcriptome, co-expression relationships may help to address these limitations by enabling the generation of unassayed gene expression data at a subcellular resolution. Here, we present CoxFormer (Co-expression pre-trained Transformer). This universal gene-embedding framework uniquely integrates literature-derived gene knowledge with data-driven co-expression networks from bulk tissues and large-scale single-cell atlases to model transcriptome-wide gene–gene relationships. CoxFormer learns robust 512-dimensional embeddings for 32,016 human genes, capturing their functional and regulatory structure. It also serves as a generative prior that is deployed through a flexible framework for transcriptome-wide spatial inference under multimodal contexts.

S055: Innovations in Regression and Change Detection for Complex Data

Cross-Semantic Transfer Learning for High-Dimensional Linear Regression

Xuejun Jiang

Southern University of Science and Technology

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.

Censored Quantile Regression With Latent Variables

Haijin He

Shenzhen University

Censored quantile regression provides valuable flexibility in assessing covariate effects on event times. We propose a censored quantile regression model with latent variables to examine the observed and latent risk factors of time-to-event of interest. The proposed model comprises two component: a confirmatory factor analysis model to characterize the latent variables through multiple observed indicators, and a censored quantile regression model that assess the effects of latent and observed variables on the survival time. We develop a hybrid procedure that combines the expectation-maximum algorithm and a smoothed and corrected estimating equation method to estimate the unknown parameters and coefficient functions. The consistency and asymptotic normality of the proposed estimators are established. We conduct simulation studies to evaluate the finite sample performance of the proposed method and apply the proposed methodology to a real-world dataset.

Bayesian Analysis of Frechet Regression

• *Hefei Liu, Yingchun Huang, Pengfei Liu, Xinyuan Song*

Yunnan University of Finance and Economics, Yunnan University of Finance and Economics, Jiangsu Normal University, The Chinese University of Hong Kong

Frechet regression extends the response variables of traditional linear regression to any metric space. This is in line with the current development direction of data science and has therefore received widely attention. At present, almost all statistical inferences about Frechet regression are based on the frequency framework, and there is no work on Bayesian analysis of Frechet regression models. This article uses Bayesian analysis to estimate and select variables for the Frechet regression models. Due to the fact that Frechet regression does not have a parametric form but only a conditional expectation form, it is not possible to give a prior on the estimated parameter. Therefore, we directly give the prior on the predicted value of the response variable. Frechet regression also does not have assumptions about error distribution, so it is not possible to obtain the likelihood of samples using traditional methods. We construct a pseudo likelihood function to describe sample information. We also improve the Bayesian information criterion for variable selection in the Frechet regression model. Simulation experiments and application examples respectively demonstrate the reliability and practicality of the method proposed in the article.

Detection of Temporal Changes in Self-Exciting Point Processes

Jiayu Huang

Central China Normal University

We study the problem of change-point detection in self-exciting point processes, where abrupt changes may occur in the dependence structure governing sequential events. Such changes are of primary interest in applications where external shocks or regime shifts alter the underlying excitation dynamics. We propose a novel procedure that formulates change-point detection as a high-dimensional optimization problem with total variation regularization, leading to a piecewise

constant estimator for the underlying model parameters and the locations of change points. The proposed approach explicitly accounts for the temporal dependence induced by self-excitation, thereby overcoming limitations of existing methods designed for independent or weakly dependent data. On the theoretical side, we establish consistency of the proposed estimator for change-point localization under general conditions, and characterize its performance in relation to the strength of dependence and the magnitude of structural changes. We conduct extensive simulation studies to assess the finite-sample performance of the proposed method and to validate the theoretical findings. We further apply the method to analyze the Ebola virus disease epidemic in the Democratic Republic of the Congo, demonstrating its ability to detect meaningful structural changes and provide interpretable insights into the evolution of the outbreak dynamics.

S073: Modern Statistical Methods for Complex Data: Hypothesis Testing, Estimation and Application

Adaptive-Gated Testing Framework for Transfer Learning: Model Structural Integrity and Functional Similarity

• *Yu Peng, Jiaqi Huang, Falong Tan, Lixing Zhu*
Beijing Normal University at Zhuhai, University of Michigan, Hunan University, Beijing Normal University at Zhuhai

The success of transfer learning is fundamentally dependent on the transferability of the source domain, a property typically evaluated by measuring the similarity between source and target regression functions. However, standard transfer methods are highly susceptible to negative transfer—either when the underlying models are mis-specified or when the functional distance between domains is excessively large.

This work addresses these vulnerabilities by proposing a unified, adaptive-to-model testing framework. This approach simultaneously verifies two critical criteria: the structural correctness of the assumed regression models and the stability of the functional distance under potential mis-specification. For the functional distance, we propose a boundary-proxy test to maintain the significance level and for model-misspecification where the predictor vectors can be sparse and ultra-high dimensional, we suggest two tests that bridge local and global smoothing tests to benefit the advantages from both methodologies. Based on adaptive gating combination of these tests, we propose an adaptive test to simultaneously handle these two testing problems. We provide a rigorous theoretical analysis of the performance of the test, demonstrating its ability to maintain the nominal significance level and its competitive power against various alternatives. Finally, numerical studies illustrate the practical effectiveness of our method in safeguarding the transfer process against both model divergence and structural misspecification, providing a reliable foundation for robust knowledge transfer.

Leveraging Big Data: Practical Strategies in Large-Scale Prospective Cohort Studies

Simin Li, Zhang Chen, Wanru Yang, Mingyun Chen, • *Xinzhi Li*
Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology, Macau University of Science and Technology

Large-scale prospective cohort studies, such as the UK Biobank (UKB) and China Kadoorie Biobank (CKB), provide unparalleled resources for advancing public health research. This report will focus on the rigorous process of big data analysis in these cohorts, including systematic data mining, covariate determination, handling of missing data, and the application of stratified and sensitivity analyses to ensure robust and reproducible findings.

We will demonstrate how advanced biostatistical strategies can be leveraged to improve cardiovascular disease (CVD) risk prediction, prevention, and diagnosis. Through the development and validation of novel cardiometabolic indices based on high-dimensional metabolomic data, we enhance traditional risk prediction models for atherosclerotic CVD. Combining machine learning with Cox proportional hazards models, these indices offer independent and incremental value for risk stratification. Furthermore, by integrating objectively measured physical activity (via accelerometers), sedentary behavior, and imaging-derived adiposity phenotypes, we elucidate the role of excises and cardiac adiposity distribution in CVD risk. Finally, stratified analyses among postmenopausal women highlight the heterogeneity of obesity and the modifying effects of menopausal hormone therapy, enabling more personalized risk assessment through nomogram-based modeling.

Collectively, these examples showcase the application of comprehensive biostatistical and methodological frameworks—from data curation to advanced modeling and validation—within large-scale cohorts. The lessons learned underscore the value of integrating multi-modal data and robust analytic strategies to inform precision cardiovascular medicine and promote effective clinical translation across diverse populations.

E-Variable-Based Framework for High-Dimensional Change Point Detection

• *Feilong Hu, Jiaqi Huang, Guanxun Li, Lixing Zhu*
Beijing normal university at Zhuhai, University of Michigan, United States, Beijing normal university at Zhuhai, Beijing normal university at Zhuhai

For a sequence of high-dimensional data, this paper proposes an

E-variable-based method to detect the structural function changes. The new method can be dimension-agnostic and sparsity-independent. It can also handle the heteroscedasticity and correlation among coordinates. We propose an Iterative DEtection Algorithm (IDEA) to enhance detection efficiency. Additionally, the coordinate-wise architecture can also simplify the construction of confidence intervals of the changes. Extensive numerical experiments validate the efficacy of the framework in the presence of high and ultrahigh dimensionality and complex heterogeneity.

Test for a Biomarker Threshold Effect in Generalized Linear Models with Application to Clinical Trials

⬩ *Bingshu Chen, Wenyu Jiang*
Queen's University, Queen's University

In a clinical trial, the responses to the new treatment may vary among patient subsets with different characteristics in a biomarker. It is often necessary to examine if there is a cutpoint for the biomarker that divides the patients into two subsets of those with more-favourable and less-favourable responses. More generally, we approach this problem as a test of homogeneity in the effects of a set of covariates in generalized linear regression models. The unknown cutpoint results in a model with nonidentifiability and a nonsmooth likelihood function to which the ordinary likelihood methods do not apply. We first use a smooth continuous function to approximate the indicator function defining the patient subsets. We then propose a penalized likelihood ratio test to overcome the model irregularities. Under the null hypothesis, we prove that the asymptotic distribution of the proposed test statistic is a mixture of chi-squared distributions. Our method is based on established asymptotic theory, is simple to use, and works in a general framework that includes logistic, Poisson, and linear regression models. In extensive simulation studies, we find that the proposed test works well in terms of size and power. We further demonstrate the use of the proposed method by applying it to clinical trial data from the Digitalis Investigation Group (DIG) on heart failure.

S088: Precision Medicine and High-Dimensional Data Inference

Optimal Treatment Regimens in Clinical Studies of Colorectal Cancer

Zicheng Lin, ⬩ *Cunjie Lin*
Renmin University of China, Renmin University of China

With the continuous advancement of medical technology, many complex diseases have become curable. In particular, for colorectal cancer—a major concern in oncology—if detected early and treated with scientific therapeutic methods, its cure rate can reach a relatively high level. Against this backdrop, in precision medicine, greater emphasis should be placed on the cure rate alongside survival time; yet most existing studies overlook the presence of cured patients. When estimating the optimal Dynamic Treatment Regime (DTR) with a cure fraction, striking a balance between the cure rate and the survival time of uncured patients has become a core issue, as both are critical to shaping optimal patient outcomes. To address this issue, this study constructs a constrained optimization framework, aiming to maximize the mean survival time of uncured individuals while ensuring a predefined cure rate. By leveraging the Lagrange function, we convert this constrained problem into an unconstrained optimization framework, facilitating more tractable analysis. A novel backward algorithm grounded in the construction of survival functions is designed to tackle the issue of partially observed survival times and cure statuses inherent in cure models. Theoretically, we derive finite-sample bounds for quantifying the discrepancy between the estimated treatment regime and the optimal DTR, using both cure rate and mean survival time of uncured patients as performance metrics. Comprehensive numerical simulations demonstrate the method's superiority across diverse scenarios, and its application to real-world colorectal cancer patient data validates its practical effectiveness, highlighting its potential to advance precision medicine treatment strategies.

Two-Way Subgroup Structure Learning for High Dimensional Clustered Survival Data

Ye He, Dongsheng Tu, Liu Liu, ⬩ *Ling Zhou*
Sichuan Normal University, Queen's University, Kingston, Canada, Sichuan Normal University, Southwestern University of Finance and Economics

Clinical investigators are interested in identification of patient subgroups that exhibit heterogeneity in clinical outcomes, as well as those with different interactions between treatment effects and multiple biomarkers. These two types of subgroups are referred to as subgroup analysis (SA) and homogeneity fusion (HF), respectively. Unlike existing methods that conduct SA or HF separately, simultaneously identifying two-way subgroup structures improves estimation and clustering accuracy, particularly in the presence of cluster heterogeneity and weak signals across high-dimensional biomarkers. In this paper, we propose a novel double-centre-augmented frailty (DCAF) model for the simultaneous identification of two-way subgroups based on clustered survival data. Instead of conventional pairwise penalties, we introduce a differentiable centre-augmented harmonic-type penalty to facilitate subgroup identification. We establish theoretical results on estimation and selection consistency. Numerical studies, including simulations and two real data examples, demonstrate the effectiveness of the proposed

DCAF method in accurately identifying two-way subgroups, as well as in selecting important biomarkers for each subgroup.

A General Distributed Privacy-Preserving Statistical Inference Framework for Massive Data

Senlin Yuan, • Xuerong Chen, Yong Zhou
Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Southwestern University of Finance and Economics, East China Normal University

In the era of massive data, privacy-preserving distributed inference plays a key role in balancing data utility, computational efficiency, and privacy protection in statistical analysis.

However, current differentially private distributed inference and federated learning methods either focus solely on parameter estimation, assume i.i.d. data, or are restricted to specific models (e.g., linear models), failing to provide comprehensive statistical inference for general models with heterogeneous data. Moreover, sensitivity calculations in existing differential privacy methods often rely on data truncation or boundedness assumptions, tailored to each model on a case-by-case basis. To address these issues, we propose a unified distributed privacy-preserving statistical inference framework suitable for both homogeneous and heterogeneous data. The framework leverages an encryption technique that integrates concepts from symmetric encryption and differential privacy. It is computationally efficient, achieves high accuracy, and eliminates the need for case-by-case sensitivity calculations. The framework is applicable to widely used models such as linear regression, logistic regression, the Cox model, and quantile regression, among others.

We also provide guidance on selecting encryption functions that can simplify both sensitivity calculations and theoretical derivations. Theoretical properties, including differential privacy guarantees, convergence rates, and the asymptotic normality of the proposed private estimators, are established. The effectiveness of our method is demonstrated through both numerical simulations and real data analyses.

Hypothesis Testing in High-Dimensional Censored-Transformation Models

Xu Liu
Shanghai University of Finance and Economics

With the rapid development of modern technologies, high-dimensional statistical inference of survival times has become increasingly important in various fields, including biostatistics and financial risk. Detecting the presence of significant high-dimensional features is crucial for accurate model interpretation and effective decision making. This paper focuses on the censored nonparametric transformation model (CNTM), whose distribution-free nature and ability to accommodate various censoring mechanisms make hypothesis testing inherently challenging. In this paper, we propose an efficient rank-based test statistic that follows an asymptotic normal distribution. The proposed test statistic allows for covariate-dependent censoring and is robust against heavy-tailed distributions and potential outliers. Furthermore, for practical purposes, we propose a new rank-based statistic to test the existence of high-dimensional features with high-dimensional control factors. Asymptotic distributions under the null hypothesis and local alternatives are established for the proposed test statistic. Numerical studies are performed to evaluate the finite sample performance of the proposed test statistics. We illustrate the proposed method with an empirical analysis of a skin cutaneous melanoma (SKCM) dataset.

S097: Recent Advances in Data-Driven Decision Making

Data-Driven Price-Setting Newsvendor Problem

Ye Lu
City University of Hong Kong

We study the price-setting newsvendor problem under incomplete demand information. We show how to use sample information to construct an ambiguity set above which we construct a robust optimization model for pricing and inventory decision. Numerical studies show that this method performs better than regression methods.

Federated Learning with Heterogeneous Data: Challenges, Algorithms, and Practical Applications

Lei Shi
Hefei University of Technology

With the increasing emphasis on data privacy and security, federated learning has become a core paradigm for distributed machine learning, enabling collaborative model training without exposing raw data. However, heterogeneous data distribution (including non-IID data, unbalanced samples, and inconsistent model structures) remains a critical bottleneck that severely degrades model performance and convergence speed in real-world scenarios.

In this talk, we will systematically analyze the root causes of performance degradation in heterogeneous federated learning, and introduce our recent research progress on personalized federated learning algorithms, including adaptive client aggregation strategies and knowledge distillation-based model alignment methods. We will also share practical deployment cases in edge computing and IoT scenarios, and discuss the future research directions of federated learning in large model era. This work aims to provide theoretical support and practical solutions for the large-scale application of federated learning in privacy-sensitive domains.

Adaptive Trial Design: Enabling Dynamic Decisions for Innovative Therapies

Ke Zhang
SCG Cell Therapy Pte. Ltd.

The clinical development of T cell-based therapies for solid tumors presents unique scientific and operational challenges arising from tumor heterogeneity, dynamic antigen expression, complex tumor microenvironments, and evolving product platforms. Conventional clinical trial designs, which rely on fixed assumptions and predetermined decision points, may be insufficient to efficiently evaluate these therapies in settings characterized by substantial biological uncertainty and rapidly emerging translational insights. Adaptive trial designs have therefore gained increasing attention as a means to enhance the efficiency, flexibility, and informativeness of early- and late-stage clinical development. This presentation examines the application of adaptive clinical trial methodologies to the development of T cell therapies for solid malignancies. We discuss key adaptive features, including model-based dose escalation, response-adaptive randomization, biomarker-driven patient selection, adaptive cohort expansion, seamless phase transitions, and pre-specified interim analyses for efficacy and futility. Particular emphasis is placed on the integration of translational biomarkers—including target antigen expression, immune monitoring, cellular kinetics, tumor microenvironment characteristics, and pharmacodynamic endpoints—to inform data-driven modifications while maintaining statistical rigor and regulatory acceptability. Through representative case studies and emerging clinical experience, we illustrate how adaptive approaches can facilitate more efficient identification of optimal dosing regimens, patient populations, and therapeutic strategies, while reducing development timelines and improving resource utilization. Furthermore, we discuss the opportunities and challenges associated with implementing adaptive designs in cell therapy trials, including operational complexity, control of type I error, manufacturing considerations, and interactions with regulatory authorities. As T cell therapies continue to evolve toward increasingly sophisticated and personalized treatment modalities, adaptive trial designs provide a powerful framework for aligning clinical development with biological learning. By enabling prospectively planned, evidence-based decision-making throughout the development process, adaptive methodologies have the potential to accelerate the translation of innovative T cell therapies into effective treatments for patients with solid tumors.

AI-Driven Personalization: Causal Learning + RL for Dynamic Treatment Decisions

Lu Wang
University of Michigan

In this talk, we present recent advances and statistical causal learning developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. We will first present a tree-based doubly robust reinforcement learning (T-RL) method, which builds a decision tree that maintains the nature of batch-mode reinforcement learning, and then a Stochastic-Tree Search method called ST-RL for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances even with a large number of covariates. In addition, we consider a common challenge with practical "restrictions" and develop a Restricted Tree-based Reinforcement Learning (RT-RL) method to address this challenge. We illustrate the method using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

S104: Recent Advances in Omics Data Analysis Techniques

Robust Statistical Methods for Analyzing Microbial Co-Abundance and Co-Occurrence Network

Yicong Mao, ⋆ *Yijuan Hu*
Peking University, Peking University

Abstract: Microbial networks characterize interdependencies (e.g., competition, commensalism, and parasitism) among taxa within communities. They can further facilitate the identification of keystone taxa, which may serve as disease biomarkers or therapeutic targets, as well as functional guilds that exhibit similar ecological roles or behaviors. However, microbiome data present unique statistical challenges, including compositionality, sparsity, and susceptibility to experimental biases and sequencing errors.

Existing approaches for co-abundance network inference typically rely on Pearson correlations computed from log-ratio transformed data, often without adequately accounting for these complexities, thereby compromising the reliability of the resulting inferences. Likewise, methods for co-occurrence network construction commonly apply Fisher's exact test to presence–absence data, without distinguishing between structural zeros and sampling zeros. To address these limitations, we propose novel and robust methodologies tailored to both settings.

Computational Modeling of Spatial Transcriptomics: From Feature Representation and Cross-Modal Prediction to Statistical Inference

Hao Wu
Shenzhen University of Advanced Technology

Spatial transcriptomics (ST) has revolutionized our understanding of tissue architecture by providing gene expression profiles within their spatial context. However, effective data interpretation remains challenging due to high dimensionality, technical noise, and cross-modal complexity. In this talk, I will present a systematic computational framework to address these challenges. First, I will introduce deep learning-based methods for gene-level representation learning, enabling robust characterization of spatial domains. Second, I will discuss our progress in cross-modal integration, specifically predicting high-resolution gene expression directly from H&E-stained histology images. Finally, I will present a rigorous statistical framework for detecting Spatially Variable Genes (SVGs) across multiple samples, facilitating the identification of consistent biological patterns. Together, these methodologies provide a robust computational toolkit for decoding the molecular organization and spatial heterogeneity of complex biological tissues.

A Bayesian Clustering Method for Spatial Omics Data Based on Factor Analysis

Yongyi Luo, Bencong Zhu, ⋆Xiaodan Fan
The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong

A typical Spatially Resolved Transcriptomics (SRT) experiment involves sectioning a three-dimensional tissue into thin slices. Each section can be considered as a technical batch. This approach introduces significant technical variations, or batch effects, between sections, which can confound biological signals and complicate integrated data analysis. To address this challenge, we propose a Bayesian method based on a Poisson factor framework that explicitly accounts for batch effects inherent in multi-section SRT data. This work provides a robust statistical foundation for the integrated analysis of SRT data across multiple tissue sections.

Keep Everything on the Same Page — A Novel Strategy for Omics Data Integrationintegration

Lixin Cheng
Shenzhen People's Hospital, the First Affiliated Hospital of Southern University of Science and Technology

The rapid development of high-throughput omics technologies has significantly advanced bioinformatics, basic medical research, and clinical diagnostics. However, most omics studies are based on single-center, small-cohort designs with limited sample sizes, making it difficult to train effective models using deep learning algorithms.

To address this challenge, we developed a qualitative analysis algorithm named Pairwise Analysis of Gene Expression (PAGE) for the integration of omics data from multiple sources. Our approach aims to: (1) design a data integration framework for multi-source, multi-platform, and multi-omics data to maximize the aggregation of informative biological signals; and (2) build generalizable computational models based on robust and reliable biological information to accurately predict clinical outcomes.

To overcome the bottleneck of small sample sizes, we have developed a series of innovative algorithms enabling cross-platform, cross-disease, and cross-age integration of biological information. Through data mapping and transfer learning, PAGE energizes the information advantage of big data into generalizable models applicable to small-data scenarios. These methods hold great potential for precision diagnosis and prognosis assessment in acute critical illnesses and cancers.

S109: Recent Advances in Statistical Designs for Early-Phase Oncology Clinical Trials

A Bayesian Pharmacokinetics-Guided Regimen Optimization with Delayed Toxicity and Survival Outcomes

Mengyi Lu, Guanfu Liu, ⋆Rongji Mu
Nanjing Medical University, ShangShanghai University of International Business and Economics, Shanghai Jiao Tong University

Delayed outcomes, such as late-onset toxicity, present substantial challenges in dose-schedule optimization during Bayesian phase I/II trials. We propose a novel B-PROSE design to identify the optimal dose-schedule regimen by integrating longitudinal pharmacokinetic (PK) data with a time-to-event survival efficacy endpoint while accounting for delayed toxicity. Our method is based on a causal pathway: dose and schedule influence PK, which in turn affects both efficacy and toxicity. We specify a joint model beginning with a Bayesian hierarchical model for the longitudinal drug concentration process. Conditional on the PK profile, we jointly model the time-to-event survival endpoint and the toxicity endpoint. The risk-benefit of each regimen is quantified using a utility function. The design continuously updates parameter estimates based on accumulating interim data and employs adaptive randomization to assign new patients to the most promising regimens. Simulation studies demonstrate that the B-PROSE design achieves desirable operating characteristics in reliably identifying the optimal regimen while safeguarding patient safety.

Great Wall: A Generalized Dose Optimization Design for Drug Combination Trials Maximizing Survival Benefit

Yan Han
Indiana University

Most phase I-II drug-combination trial designs assume that selecting the optimal dose combination based on early outcomes

will also lead to maximum long-term survival benefits. However, this assumption is often violated in many clinical studies, generally due to high rates of relapse following the initial response. To address this problem, we propose the Great Wall design, a general dose optimization design for drug-combination trials. The Great Wall design employs a "divide-and-conquer" algorithm to address the issue of partial order of toxicity and uses early outcomes to eliminate dose combinations that are excessively toxic or less efficacious. It utilizes a dose randomization approach to construct a candidate set of the promising dose combinations balancing the toxicity and early efficacy outcomes. The patients assigned to the candidate set are followed to collect the survival outcomes and the final optimal dose combination is then selected to maximize the survival benefit. The simulation studies confirm the desirable operating characteristics of the Great Wall design under various clinical settings. R codes are also provided to facilitate the application. The Great Wall design is modular and practically useful in settings where investigators plan to follow patients long enough to assess survival outcomes.

Precision Generalized Phase I-II Designs

Yong Zang

Indiana University School of Medicine

A new family of precision Bayesian dose optimization designs, PGen I-II, based on early efficacy, early toxicity, and long-term time to treatment failure is proposed. A PGen I-II design refines a Gen I-II design by accounting for patient heterogeneity characterized by subgroups that may be defined by prognostic levels, disease subtypes, or biomarker categories. The design makes subgroup-specific decisions, which may be to drop an unacceptably toxic or inefficacious dose, randomize patients among acceptable doses, or identify a best dose in terms of treatment success defined in terms of time to failure over long-term follow-up. A piecewise exponential distribution for failure time is assumed, including subgroup-specific effects of dose, response, and toxicity. Latent variables are used to adaptively cluster subgroups found to have similar dose-outcome distributions, with the model simplified to borrow strength between subgroups in the same cluster. Guidelines and user-friendly computer software for implementing the design are provided. A simulation study is reported that shows the PGen I-II design is superior to similarly structured designs that either assume patient homogeneity or conduct separate trials within subgroups.

S114: Recent Advances on the Analysis of Complex Data

Novel Empirical Likelihood Method for the Cumulative Hazard Ratio Under Stratified Cox Models

Dazhi Zhao, ⋆ *Yichuan Zhao*

Georgia State University, Georgia State University

Evaluating the treatment effect is a crucial topic in clinical studies. Nowadays, the ratio of cumulative hazards is often applied to accomplish this task, especially when those hazards may be nonproportional. The stratified Cox proportional hazards model, as an important extension of the classical Cox model, has the ability to flexibly handle nonproportional hazards. In this article, we propose a novel empirical likelihood method to construct the confidence interval for cumulative hazard ratio under the stratified Cox model. The large sample properties of the proposed profile empirical likelihood ratio statistic are investigated, and the finite sample properties of the empirical likelihood-based estimators under some different situations are explored in simulation studies. The proposed method was finally applied to perform statistical analysis on a real-world dataset on the survival experience of patients with heart failure.

Variable Selection with Broken Adaptive Ridge Regression for Interval-Censored Competing Risks Data

Fatemeh Mahmoudi, Chenxi Li, Kaida Cai, ⋆ *Xuewen Lu*

Mount Royal University, Michigan State University, Southeast University, University of Calgary

Competing risks data refer to situations in which the occurrence of one event precludes the occurrence of other events, resulting in multiple mutually exclusive events. This data type is commonly encountered in medical research and clinical trials, where it is used to explore the interplay among events and to inform decision-making in fields such as healthcare and epidemiology. We develop a penalized variable selection procedure to handle such complex data in an interval-censored setting. We consider a broad class of semiparametric transformation regression models, including popular models such as proportional and non-proportional hazards models. To promote sparsity and select variables specific to each event, we employ the broken adaptive ridge (BAR) penalty. This approach allows us to simultaneously select important risk factors and estimate their effects for each event under investigation. We establish the oracle property of the BAR procedure and evaluate its performance through simulation studies. The proposed method is applied to a real-life HIV cohort dataset, further validating its applicability in practice.

Estimation of Stratified Semiparametric Transformation Models Under Interval Censoring with Auxiliary Information

⋆ *Yinghao Pan, Yanqing Sun, Xi Ning, Peter Gilbert*

University of North Carolina at Charlotte, University of North Carolina at Charlotte, Colby College, Fred Hutchinson Cancer Center

Interval-censored data are prevalent in biomedical, clinical, and epidemiological studies. In this paper, we investigate how to more efficiently estimate stratified semiparametric transformation models under interval censoring when auxiliary failure time information is available. Specifically, we propose a computationally efficient and reliable EM algorithm in which the weights incorporate the auxiliary information. Extensive simulation studies, along with an application to HIV prevention trials, demonstrate the superiority of the proposed method.

A Factor-Augmented Regression Approach to Learning Bivariate Interval-Censored Data

Shuwei Li
Guangzhou University

Comorbidity, such as the coexistence of hypertension and diabetes, poses a major challenge in modern medicine, as it reflects the systemic and holistic nature of human health. Understanding the latent risk factors underlying comorbidity is essential for the prevention and treatment of diseases from a more comprehensive and integrated perspective. In real-life disease surveillance, comorbidity occurrence is often monitored through periodic examinations, leading to the frequent emergence of bivariate interval-censored data. In this article, we propose a generalized exploratory factor analysis model within a likelihood-based framework and integrate it with a class of time-to-event frailty models. This joint modeling framework enables the extraction of latent factors from various types of covariates while simultaneously assessing their effects on bivariate interval-censored failure events. We focus on nonparametric maximum likelihood estimation and develop an expectation-maximization algorithm to achieve reliable and efficient implementation. Sufficient conditions for model identifiability are established, and the asymptotic properties of the proposed estimators are derived. The proposed method demonstrates satisfactory performance in simulation studies and is further applied to a real-world data analysis of Type 2 diabetes, showing its desirable abilities in addressing the multicollinearity issue and uncovering the latent covariate structure.

S126: Statistical and Computational Methods for Omics and Biological Discovery

Structured and Scalable AI for High-Dimensional Biological Data: Beyond Lasso Under Correlation

Xiaoqing Huang, Aolong Li, ⋆ *Yijie Wang*
Indiana University Bloomington, Indiana University Bloomington

High-dimensional biological data pose fundamental challenges for artificial intelligence due to strong feature correlations, limited sample sizes, and the need to balance predictive and explanatory power. Traditional sparsity-inducing methods, such as Lasso, often fail under high correlation, leading to unstable feature selection and reduced interpretability. In this work, we present a scalable optimization framework based on Boolean convex programming to address these limitations. Our approach reformulates subset selection using a tight convex relaxation with theoretical guarantees, enabling efficient optimization via projected quasi-Newton methods. We further extend this framework to structured sparsity settings, including group, overlapping, and graph-based regularization. Empirical evaluations on genome-wide association studies (GWAS) demonstrate improved scalability, robustness to correlation, and enhanced feature selection accuracy compared to existing methods. This work provides a unified pathway from classical sparsity models to modern structured AI methods for biological data analysis.

Model-Based Multifacet Clustering with High-Dimensional Omics Applications

⋆ *Wei Zong, Danyang Li, Marianne Seney, Colleen Mcclung, George C. Tseng*
The Chinese University of Hong Kong, Shenzhen, University of Pittsburgh, University of Pittsburgh, University of Pittsburgh, University of Pittsburgh

High-dimensional omics data often contain intricate and multifaceted information, resulting in the coexistence of multiple plausible sample partitions based on different subsets of selected features. Conventional clustering methods typically yield only one clustering solution, limiting their capacity to fully capture all facets of cluster structures in high-dimensional data. To address this challenge, we propose a model-based multifacet clustering (MFClust) method based on a mixture of Gaussian mixture models, where the former mixture achieves facet assignment for gene features and the latter mixture determines cluster assignment of samples. We demonstrate superior facet and cluster assignment accuracy of MFClust through simulation studies. The proposed method is applied to three transcriptomic applications from postmortem brain and lung disease studies. The result captures multifacet clustering structures associated with critical clinical variables and provides intriguing biological insights for further hypothesis generation and discovery.

Covariate-Informed Multiple Testing via the Conditional Local False Discovery Rate, with an Application to a Melanoma Study

Ziyi Wang, Satabdi Saha, Christine Peterson, ⋆ *Yushu Shi*
University of Texas Health, University of Texas MD Anderson Cancer Center, Rice University, Weill Cornell Medicine

We develop chai (Conditional Hypothesis testing using Auxiliary Information), a covariate-informed multiple testing procedure based on the conditional local false discovery rate (clfdr). Motivated by high-dimensional microbiome studies, where rich side information such as phylogenetic structure, functional pathways, or metabolite profiles is often available, chai aims to increase power while controlling the false discovery rate (FDR). Our method models the joint distribution of test statistics and covariates with a flexible Gaussian mixture, derives the conditional distribution of the z-statistic given its covariates, and combines this with a deconvolution step to obtain hypothesis-specific null probabilities. A leave-one-out strategy is used to avoid double-dipping when estimating mixture parameters, yielding better-calibrated clfdr estimates, and hypotheses are selected by thresholding the ordered clfdr values so that the average posterior null probability is controlled at the target FDR level. Extensive simulations show that chai consistently attains higher power than existing methods while maintaining empirical FDR near the nominal level. An application to a melanoma dataset focusing on anti-PD-1 response with literature-derived covariates demonstrates that chai yields more discoveries that are biologically interpretable and supported by prior evidence, especially when multiple correlated sources of auxiliary information are available.

S128: Statistical Foundations for Machine Learning and AI

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

Zihan Dong, Xiaotian Hou, ⋆Ruijia Wu, Linjun Zhang

Rutgers University, University of Pennsylvania, Shanghai Jiao Tong University, Rutgers University

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 (\method), 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 \method~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. Simulations and real-data analysis demonstrate the practical benefits and superior performance of our proposed method.

Data Selection for LLM: Evolving from Closed to Open

Jun Shu

Xi'an Jiaotong University

The performance of large language models (LLMs) is critically shaped by the distributional composition of multi-domain pre-training data. While early approaches relied on manual heuristics, such strategies fail to capture complex cross-domain interactions as data scales. A dominant paradigm models the relationship between domain weights and validation loss via proxy functions and then optimizes the weights accordingly. However, these approaches depend on strong structural assumptions (e.g., rank invariance or scaling laws), which are often violated in practice, leading to non-negligible estimation bias.

An alternative line of work directly optimizes domain weights from data, but suffers from unstable optimization trajectories and high computational cost. To address these challenges, we propose a Bayesian domain weighting framework that infers domain weights from a Dirichlet distribution, with Gamma priors learned from observations. Experimental results demonstrate that the proposed method achieves stable and efficient optimization, identifies high-quality domain mixtures, and significantly reduces data requirements compared to search-based approaches, thereby enabling scalable domain weighting for large-scale LLM training.

Spatially Randomized Designs Can Enhance Policy Evaluation

Ying Yang

Fudan University

This article studies the benefits of using spatially randomized experimental designs which partition the experimental area into distinct, non-overlapping units with treatments assigned randomly. Such designs offer improved policy evaluation in online experiments by providing more precise policy value estimators and more effective testing algorithms than traditional global designs, which apply the same treatment across all units simultaneously. We examine both parametric and nonparametric methods for estimating and inferring policy values based on the spatially randomized designs. Our analysis includes evaluating the mean squared error of the treatment effect estimator and the statistical power of the associated tests. Additionally, we extend our findings to the dynamic setting with spatio-temporal dependencies, where treatments are allocated sequentially over time, and account for potential temporal carryover effects. Our

theoretical insights are supported by comprehensive numerical experiments.

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

Chendi Wang
Xiamen University

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.

S050: Frontiers of Transfer Learning, Data Fusion, and Adaptive Evidence Generation: Theory, Methods, and Practice

Targeted Learning via Probabilistic Subpopulation Matching

• *Xiaokang Liu, Jie Hu, Naimin Jing, Yang Ning, Cheng Yong Tang, Runze Li, Yong Chen*

University of Missouri, University of Pennsylvania, Merck & Co., Inc, Cornell University, Temple University, Penn State University, University of Pennsylvania

In biomedical research, to obtain more accurate prediction results from a target study, leveraging information from multiple similar source studies is proved to be useful. However, in many biomedical applications based on real-world data, populations under consideration in different studies, e.g., clinical sites, can be heterogeneous, leading to challenges in properly borrowing information towards the target study. The state of art methods are typically based on study-level matching to identify source studies that are similar to the target study, whilst samples from source studies that significantly differ from the target study will all be dropped at the study level, which can lead to substantial loss of information. We consider a general situation where all studies are sampled from a super-population composed of distinct subpopulations, and propose a novel framework of targeted learning via subpopulation matching. In contrast to the existing study-level matching methods, measuring similarities between subpopulations can effectively decompose both within- and between-study heterogeneity, allowing incorporation of information from all source studies without dropping any samples as in the existing methods. We devise the proposed framework as a two-step procedure, where a finite mixture model is first fitted jointly across all studies to provide subject-wise probabilistic subpopulation information, followed by a step of within-subpopulation information transferring from source studies to the target study for each identified subpopulation. We establish the non-asymptotic properties of our estimator and demonstrate the ability of our method to improve prediction at the target study via simulation studies.

Efficient Semiparametric Inference for Distributed Data with Blockwise Missingness

Jingyue Huang, • *Huiyuan Wang, Yuqing Lei, Yong Chen*

University of Pennsylvania, University of Pennsylvania, University of Pennsylvania, University of Pennsylvania

We consider statistical inference for a finite-dimensional parameter in a regular semiparametric model under a distributed setting with blockwise missingness, where entire blocks of variables are unavailable at certain sites and sharing individual-level data is not allowed. To improve efficiency of the internal study, we propose a class of augmented one-step estimators that incorporate information from external sites through "transfer functions." The proposed approach has several advantages. First, it is communication-efficient, requiring only one-round communication of summary-level statistics. Second, it satisfies a do-no-harm property in the sense that the augmented estimator is no less efficient than the original one based solely on the internal data. Third, it is statistically optimal, achieving the semiparametric efficiency bound when the transfer function is appropriately estimated from data. Finally, it is scalable, remaining asymptotically normal even when the number of external sites and the data dimension grow exponentially with the internal sample size. Simulation and real-world studies confirm both the statistical efficiency and computational feasibility of our method in distributed settings.

Surrogate-Powered Inference

• *Jianmin Chen, Huiyuan Wang, Thomas Lumley, Xiaowu Dai, Yong Chen*

Xiamen University,University of Pennsylvania, University of Auckland, University of California, Los Angeles, University of

Pennsylvania

High-quality labeled data are essential for reliable statistical inference, but are often limited by validation costs. While surrogate labels provide cost-effective alternatives, their noise can introduce non-negligible bias. To address this challenge, we propose the surrogate-powered inference (SPI) toolbox, a unified framework that leverages both the validity of high-quality labels and the abundance of surrogates to enable reliable statistical inference. SPI comprises three progressively enhanced versions. Base-SPI integrates validated labels and surrogates through augmentation to improve estimation efficiency. SPI+ incorporates regularized regression to safely handle multiple surrogates, preventing performance degradation due to error accumulation. SPI++ further optimizes efficiency under limited validation budgets through an adaptive, multiwave labeling procedure that prioritizes informative subjects for labeling. Compared to traditional methods, SPI substantially reduces the estimation error and increases the power in risk factor identification. These results demonstrate the value of SPI in improving the reproducibility. Theoretical guarantees and extensive simulation studies further illustrate the properties of our approach.

A Unified Information–Time Paradigm for Single-Arm Time-to-Event Trials: The Window-Cohort with Calibrated Follow-Up Requirement

Haitao Pan

St. Jude Children's Research Hospital

Single-arm phase II trials with time-to-event endpoints are commonly designed using event-driven or enrollment-driven interim rules, which control statistical information or calendar time separately but not jointly. This limitation is particularly problematic in rare-disease and small-population settings: event-driven designs may lead to prolonged and unpredictable delays when events are sparse, whereas enrollment-driven designs may trigger interim decisions based on immature follow-up. In addition, the effective information at interim depends on the estimand of interest, such as milestone survival at a fixed time point or treatment effects under proportional hazards.

We propose a unified framework that jointly controls information and time by aligning interim timing with the estimand-specific information structure. The proposed Window-Cohort with Calibrated Follow-Up Requirement (WCR) design triggers interim analysis at the enrollment of the Nw-th patient, followed by a calibrated post-lock follow-up window X. This construction ensures a predictable and interpretable information horizon at interim while allowing continued accrual to contribute to the final analysis. We distinguish restricted and unrestricted follow-up regimes corresponding to landmark-based and proportional hazards–based estimands and characterize their implications for the effective data horizon.

Design parameters, including Nw, X, total sample size, and decision thresholds, are selected via constrained optimization to control type I error and power while optimizing efficiency metrics such as expected sample size and calendar time. Simulation studies demonstrate that the proposed framework preserves nominal operating characteristics and improves calendar efficiency and stability compared to conventional designs.

The methodology is implemented in the open-source R package WCRBayesDesign, facilitating reproducible calibration and transparent reporting in clinical trial development.

S136: Statistical Learning and Modeling for Complex Data

Utilizing Large Models in Biomedical Data Analysis: Some Case Studies

Yinglei Lai

George Washington University

In this talk, based on some case studies, we introduce our recent work on the use of large models in biomedical data analysis. The first case study is on the correction of batch-effects. The related methods have been widely developed for RNA-sequencing data. However, there is still a lack of methods for preserving the numerical order of data. We have recently achieved this based on a monotonic deep learning framework. The second case study is on the information extraction from clinical notes. Identifying patients with critical illness in emergency departments (EDs) is an ongoing challenge, partly due to the limited information available at the time of admission. The utilization of large language models for analyzing clinical notes has not been extensively investigated. We have recently achieved this to improve the severity assessment based on ED patient complaints and anamneses in their clinical notes. More case studies will also be presented for this talk when time permits.

Fuzzy Join of Data for Statistical Learning

⬩*Donghui Yan, Yingjie Wang, Haiping Xu, Chenyu Li, Ryan Berry, Hongyu Zhao*

UMass Dartmouth, UMass Dartmouth, UMass Dartmouth, Columbia University, UMass Dartmouth, Yale University

With the emergence of big data and distributed computing, it is increasingly often that the dataset of interest has features shared by some external auxiliary data. Inspired by the join operation in relational database, we propose an efficient algorithm—fuzzy join—that would enhance statistical learning by levering the auxiliary data. Fuzzy join extends the scope of relational database join by allowing it to be performed on non-index key columns and

also allows non-exact match between rows from different datasets. It can be used to bring in extra information along features present in the auxiliary but not the original data thus enhanced statistical learning is expected. Experiments on synthesized and benchmark datasets demonstrate that fuzzy join is feasible and can potentially help improve the performance of a range of statistical learning tasks including classification, clustering and anomaly detection etc. Some theoretical insights have been developed to support the empirical success of fuzzy join.

Structural Learning of Arbitrary Graphs under Potts Model Framework

Xinran Bian
Sun Yat-Sen University

Consider the problem of learning the graph underlying a Potts model from a collection of i.i.d. samples generated from the model. In this paper, we describe a greedy algorithm that is able to recover the structure of a Potts model on an arbitrary bounded-degree graph in time on the order of p2. We do not require any additional assumptions on the parameters except what is necessary for the identifiability of the Potts model.

Our algorithm leverages conditional influence measures to iteratively identify node neighborhoods and achieves exact structure recovery with high probability, provided the sample size satisfies theoretical bounds. We establish rigorous guarantees on the sample complexity and runtime, demonstrating that our method performs efficiently even in high-dimensional settings. Empirical evaluations on both synthetic and real datasets (Visual Coding Neuropixels, CuMiDa, and Seattle Alzheimer's dataset) show that our approach outperforms existing methods, such as the Peter-Clark (PC) and the Graphical LASSO (GLASSO) algorithm, particularly in scenarios involving multi-state, non-Gaussian data. This work advances the understanding of structure learning for general MRFs, bridging the gap between Ising and Potts models, and offers a robust tool for network recovery in complex non-binary systems such as neural networks and gene-gene interactions.

Covariate-Adjusted Win Statistics in Randomized Clinical Trials with Ordinal Outcomes

Zhiqiang Cao
Shenzhen Technology University

In this article, we represent the average treatment effect for ordinal outcomes via pairwise outcome comparisons captured through win estimands, such as the win ratio and win difference. To perform covariate adjustment, we develop propensity score weighting estimators, including inverse probability weighting (IPW) and overlap weighting (OW), tailored to estimating win estimands, and further investigate augmented weighting estimators with an outcome regression component. Leveraging the theory of U-statistics, we establish the asymptotic theory for all estimators, and derive closed-form variance estimators to support inference. We compare our estimators via simulations, illustrate our proposed methods with the ORCHID trial, and implement the proposed methods in the winPSW R package.

S147: Study Design and Joint Modeling with Longitudinal and Survival Data

Longitudinal Variable Selection Under Joint Modeling of Longitudinal and Binary Data

◆ *Fangfang Wang, Lihui Zhao*
Tianjin University of Finance and Economics, Northwestern University

In this article, we propose a longitudinal variable selection method for the joint modeling of multivariate longitudinal measurements and binary outcomes. The longitudinal process is characterized by a multivariate linear mixed-effects model, while the binary outcome is modeled using a logistic regression. We employ an Expecta- tion/Conditional Maximization Either (ECME) algorithm to simultaneously perform variable selection and estimation for both longitudinal variables and fixed effects, which incorporates a pseudo-likelihood approach for the binary component to facilitate analytic solutions. Random effects are estimated via best linear unbiased prediction (BLUP) method. For variable selection, a Group Lasso penalty is imposed on the coefficients of the random intercept and random slope in the binary submodel to identify influential longitudinal variables. Simultaneously, Lasso penalties are applied to the fixed effects in both submodels to select relevant covariates. Simulation studies demonstrate that the proposed method achieves excellent performance in both parameter estimation and variable selection.

Parsimonious Joint Model of Survival Outcome and Multiple Longitudinal Risk Factors

◆ *Zhiyang Zhou, Lihui Zhao*

University of Florida

Dynamic prediction updates an individual's risk assessment for a particular outcome by incorporating new information over time. The primary challenge lies in accurately modeling the complex interaction between longitudinal and survival outcomes. The shared-random-effects joint model, a widely used approach for dynamic prediction, simultaneously fits submodels for longitudinal and survival outcomes. However, as the number of longitudinal outcomes increases, the number of unknown parameters grows rapidly, leading to substantial computational burden. This complexity may also degrade predictive accuracy due to approximation errors arising from an intricate likelihood

function. To address these challenges, we draw inspiration from exploratory factor analysis (EFA) and propose the parsimonious joint model (pJM) that represents random effects using a small set of mutually independent latent factors. Unlike EFA, pJM enjoys rigorously established identifiability while reducing the number of model parameters. In addition, we employ variational inference to further improve computational efficiency in model fitting. Numerical studies validate pJM, demonstrating its competitive predictive accuracy.

Integrating Intensive Mobile Monitoring Data for Dynamic Prediction of Healthcare Utilization via Landmark Models and FPCA

• *Qing Yang, Anushka Palipana, Wenbin Lu, Ryan Shaw, Matt Crowly*
Duke University, Duke University, North Carolina State University, Duke University, Duke University

Background:
Mobile health (mHealth) technologies enable continuous monitoring of patient behaviors and physiology, offering opportunities to improve prediction of healthcare utilization beyond traditional clinical and demographic data. The EXTEND study is a pragmatic, randomized trial evaluating comprehensive telehealth for patients with uncontrolled type 2 diabetes and hypertension. While the primary EXTEND trial focuses on intervention effectiveness, this analysis uses the same dataset for an exploratory objective examining the added value of device-derived data (e.g., blood glucose, blood pressure, weight, and step count) in prediction models.

Objective:

To assess whether integrating mobile monitoring data with baseline demographic characteristics and longitudinal clinical biomarkers (HbA1c etc.) improves prediction of healthcare utilization outcomes, including hospitalizations and emergency department visits, over 24 months in the EXTEND study cohort.

Methods:
We applied a landmark modeling framework to 24-month longitudinal data from the EXTEND study. Two baseline models were developed: (1) a model including key demographic variables and the most recent values of longitudinal clinical biomarkers, and (2) a model incorporating subject-specific biomarker estimates obtained from linear mixed-effects models at prespecified landmark time points over follow-up. Frailty Cox proportional hazards models were used to account for recurrent utilization events. To incorporate high-frequency mobile monitoring data, functional principal component analysis (FPCA) was used to summarize intensive longitudinal device measurements over 24 months. The resulting functional features were included in the second-stage models, with regularization-based variable selection applied to identify predictive features. Predictive performance was compared between models with and without mobile device data to quantify the incremental value of digital monitoring information.

Results:

Preliminary analyses demonstrate the feasibility of integrating longitudinal clinical and device data within a dynamic prediction framework. Ongoing analyses are evaluating the extent to which device-derived features improve prediction of healthcare utilization outcomes.

Regression-Based Modeling of Spearman's Rho for Longitudinal Metabolomics and Mental Wellness in Breast Cancer Patients

Tuo Lin
University of Florida

Chemotherapy in breast cancer (BC) can substantially affect mental wellness. Advances in metabolomics enable comprehensive profiling of metabolic changes over time during and after treatment, offering insights into biological mechanisms linking chemotherapy to mental health outcomes. To study the association between metabolite profiles and mental wellness, correlation-based analyses are particularly useful. Spearman's rho is a widely used correlation measure and popular alternative to Pearson's correlation, since it also applies to non-linear association between variables. However, existing methods are not designed for longitudinal data and do not allow for covariate adjustments. In this work, we propose a novel regression-based framework grounded in a class of semiparametric models, the functional response models, to extend this popular correlation measure to longitudinal settings with missing data under the missing at random assumption. This framework facilitates inferences about temporal changes in correlations over time and association of explanatory variables for such changes. We use simulation studies to evaluate performance of the approach with moderate sample sizes. We apply the approach to a one-year longitudinal substudy of the EPIGEN study to examine the longitudinal association between metabolite profiles and mental wellness in BC patients undergoing chemotherapy. The identified metabolites may serve as candidates for future in-depth bioinformatics analyses and translational investigations.

S154: When Large Language Model Meets Statistics

Evaluating Uncertainty of Large Language Models via Distributional and Structural Estimation

Pengkun Yang
Tsinghua University

Evaluating the uncertainty of large language models is essential

for reliable deployment, yet remains challenging due to complex output distributions, hallucinated reasoning, and inconsistent comparative judgments. We propose a distributional estimation approach for modeling large-scale tokenized outputs via mixtures over unlabeled histograms, establishing optimal convergence guarantees and enabling efficient uncertainty quantification for downstream functionals. Additionally, we introduce a structural reasoning model that represents chain-of-thought outputs as directed acyclic graphs, and developing a prompt perturbation approach that identifies and filters structurally inconsistent comparison patterns. Experiments on benchmark datasets and multiple model families demonstrate improved hallucination detection, ranking consistency, and estimation accuracy, highlighting the benefits of integrating statistical and structural approaches for principled LLM uncertainty evaluation.

When Kernel Ridge Regression Meets Large Language Models

Xin Bing, ⋆Xin He, Chao Wang
University of Toronto, Shanghai University of Finance and Economics, University of Toronto

Kernel methods, particularly kernel ridge regression (KRR), are time-proven, powerful nonparametric regression techniques known for their rich capacity, analytical simplicity, and computational tractability. The analysis of their predictive performance has received continuous attention for more than two decades. However, in many modern regression problems where the feature inputs used in KRR cannot be directly observed and must instead be inferred from other measurements, the theoretical foundations of KRR remain largely unexplored. In this talk, we introduce a novel approach for analyzing KRR with latent feature inputs, and we also apply our general results to some applications. Our theoretical findings are further corroborated by simulation studies and real-data analyses using pretrained LLM embeddings for the downstream prediction task.

Efficient Reasoning with Statistical Guarantees for Large Language Models

⋆*Hao Zeng, Jianguo Huang, Bingyi Jing, Hongxin Wei, Bo An*
Southern University of Science and Technology

Large reasoning models (LRMs) excel at complex tasks but are computationally expensive to run. While dynamically switching between high-cost “thinking” and low-cost “non-thinking” modes can improve efficiency, current methods lack guarantees and can introduce errors. To address this, we propose PAC Reasoning, a framework that provides statistical guarantees on performance loss when making such switches. Our method calculates a confidence bound for the potential performance loss based on model uncertainty and uses this to set a switching threshold. This ensures the performance loss remains within a user-specified tolerance. Theoretical analysis proves our approach bounds the performance loss in a distribution-free manner. Experiments on standard benchmarks demonstrate that PAC Reasoning effectively reduces computational costs while reliably controlling the performance loss as specified by the user.

AlphaPareto: Formulaic Alpha Discovery with LLM-Guided Multi-Objective Reinforcement Learning

Yingbo Zhao, Zeyu Yang, ⋆ Zhoufan Zhu
Xiamen University, Xiamen University, Xiamen University

Formulaic alpha discovery is a fundamental problem in quantitative trading, yet identifying alphas that work well together remains challenging. Recent reinforcement learning (RL) methods formulate this task as a Markov decision process (MDP), but two important issues remain unresolved. First, as the alpha pool evolves, the reward function changes accordingly, making the MDP inherently non-stationary. Second, most existing methods optimize a single objective, typically predictive power, while ignoring other important properties of a high-quality alpha pool. Motivated by these challenges, we propose AlphaPareto, an RL method for formulaic alpha discovery. To mitigate non-stationarity, AlphaPareto augments the state to include both the alpha under construction and the current alpha pool, and applies a large language model (LLM) to encode the pool. This design allows the agent to adapt to the evolving search environment. To overcome the limitation of single-objective reward design, AlphaPareto replaces the scalar reward with a multi-objective vector-valued reward that simultaneously captures predictive power, temporal stability, perturbation robustness, and diversity, and optimizes these objectives through a Pareto-regularized learning procedure. Empirical applications to real-world datasets show that our AlphaPareto method significantly outperforms its competitors.

S005: Advanced Statistical Methods for Complex Survival Data

Distributed Censored Quantile Regression: Convolution Smoothing and Communication Efficiency

Ying Huang, Xiaoqi Jiao, ⋆ Huijuan Ma, Baoxue Zhang
East China Normal University

Censored quantle regression (CQR) has become a popular framework for analyzing survival outcomes, yet conventional estimation procedures facecomputational challenge when scaled to massive datasets. In such modern large-scale settings, individual-level survival data cannot be freelyaccessed or centrally pooled due to privacy concems and storage limitations, rendering distributed computation increasingly indispensable. The nondifferentiability of the CQR loss further complicates optimzation, making exsting methods computationaly expensive and poorly suited fordistributed data structures. To

overcome these challenges, we propose a communication-efficient distributed CQR framework that leveragesonvolution smoothing to construct a globaly smooth and convex objective, thereby enabling quasi-Newton optimization in large- scale distributednvironments. The proposed approach effectively balances computational scalability and communication efficiency. We establish the Bahadurepresentation and asymptotic properties of the convolution-smoothed CQR estimator on a single machine, and further derive the convergence rateof the distributed quasi-Newton updates. Extensive simulations and real-world applications demonstrate that the proposed method achievesstatistical acuracy comparable to single- machine ful-sample algorithms, while outperforming existing distributed CQR methods in bothcomputational and communication efficiency.

Regression Analysis of Dependent Current Status Data with Left-Truncation Under Linear Transformation Model

Mengyue Zhang, Shishun Zhao, ⋆Da Xu, Tao Hu, Jianguo Sun
Jilin University, Jilin University, Northeast Normal University, Capital Normal University, Southern University of Science and Technology

The paper discusses the regression analysis of current status data, which is common in various fields such as tumorigenic research and demographic studies. Analyzing this type of data poses a significant challenge and has recently gained considerable interest. Furthermore, the authors consider an even more difficult scenario where, apart from censoring, one also faces left-truncation and informative censoring, meaning that there is a potential correlation between the examination time and the failure time of interest. The authors propose a sieve maximum likelihood estimation (MLE) method and in the proposed method for inference, a copula-based procedure is applied to depict the informative censoring. Additionally, the authors utilise the splines to estimate the unknown nonparametric functions in the model, and the asymptotic properties of the proposed estimator are established. The simulation results indicate that the developed approach is effective in practice, and it has been successfully applied to a set of real data.

Dynamic Survival Prediction Based on Multiple Intermediate Events

⋆*Tonghui Yu, Liming Xiang*
Harbin Institute of Technology, China, Nanyang Technological University, Singapore

In many follow-up studies, the onset of intermediate events is crucial for predicting the overall terminal event. However, these intermediate events may be potentially associated, exhibit partial-order relations and suffer from informative censoring caused by the terminal event, posing challenges for both modelling and inference. We propose a novel joint modeling framework to build dependence among multiple correlated marginal components. Under this framework, we develop a novel algorithm and a renewable risk assessment method for dynamic survival prediction, leveraging the available information of intermediate events and the maximum follow-up time.

Online Updating and Heterogeneity Detection for Additive Hazard Models with Streaming Survival Data

Lina Wang, Zhanhang Gao, ⋆Bing Wang
Anhui University, Anhui University, Anhui University
In the era of big data, survival data are often collected sequentially in the form of data streams, posing substantial computational and storage challenges to conventional batch-processing methods. In this paper, we propose an efficient online updating framework for the additive hazard model to analyze streaming survival data. Unlike traditional approaches that require repeated access to the full dataset, the proposed method dynamically updates regression coefficients and baseline hazard functions using only newly arriving data batches and a set of historical summary statistics, thereby substantially reducing computational cost and memory requirements. Our framework provides closed-form online variance estimators for both regression coefficients and baseline hazards, enabling valid statistical inference in a streaming environment. To accommodate potential heterogeneity across data batches, we further incorporate an adaptive Lasso procedure to detect and quantify batch-to-batch changes in model parameters. Extensive simulation studies demonstrate that the proposed estimators achieve statistical accuracy comparable to full-data analyses while offering significant computational efficiency gains. An application to the SEER lung cancer dataset further illustrates the practical utility of the proposed method in large-scale survival analysis.

S021: Advances in Statistical Modeling for Causal Inference and High-dimensional Complex Data

Estimating Conditional Complier Quantile Treatment Effect Via Stratified Quantile Regression

Huijuan Ma, ⋆Mengjiao Peng, Jing Qin

East China Normal University, East China Normal University, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Understanding the causal effect of a treatment in randomized experiments with non- compliance is of fundamental interest in many domains. Within the instrumental variable (IV) framework, the causal treatment effect can only be reliably assessed for compliers, as they are the only subpopulation whose treatment assignment is influenced by the instrument. In this article, we

study the conditional complier quantile treatment effect based on individual characteristics through stratified quantile regression models for compliers with and without treatment, which are flexible in capturing the interaction between treatment and covariates and include the past unified model as a special case. We introduce a tuning parameter-free method that directly utilizes the mixture structure in the compiler problem, departing from past approaches that relied on minimizing a weighted check function with nonparametric method- estimated weights. A novel iterated algorithm is proposed to solve discontinuous equations that involve unknown parameters in a complicated way. The consistency and asymptotic normality of the proposed estimators are established. Numerical results, including extensive simulation studies and real data analysis of the Oregon health insurance experiment and a job training study, show the practical utility of the proposed approach.

Optimal Dynamic Treatment Regimes for High-Dimensional Accelerated Failure Time Model

Shijie Zhao, ⋆ *Wei Zhao*

Shandong University, Shandong University

The increasing prevalence of high-dimensional covariates presents a significant challenge to precision medicine. To overcome this, we propose a novel multi-stage optimal dynamic treatment regime method specifically designed for high-dimensional accelerated failure time models. Our approach, which aims to maximize individual survival time, employs backward induction with counterfactual survival times as the optimization objective. Parameter estimation leverages weighted counterfactual survival times to account for censoring, followed by the Slope Determination by Analysis of Residuals (SDAR) algorithm. Theoretical analysis guarantees exponential convergence with non-asymptotic error bounds under mild conditions. Simulation studies demonstrate a substantial improvement in survival time estimation, particularly under high censoring rates, compared to existing methods. We illustrate the practical utility of our method by applying it to a sepsis dataset, revealing clinically meaningful treatment recommendations.

Variable Selection for Mixed Panel Count Data Using Gaussian Seamless-L0 Penalty

⋆ *Lei Ge, Tao Hu, Shurong Zheng, Jianguo Sun*

Northeast Normal University, Capital Normal University, Northeast Normal University, Southern University of Science and Technology

This paper considers variable selection for mixed panel count data, which frequently occur in longitudinal studies and whose analysis is quite challenging due to their complex data structures. %when integrating both panel count and panel binary components.

For the problem, we propose a penalized likelihood procedure with the use of Gaussian Seamless-L0 penalty under a proportional mean model. For its implementation, a computationally efficient EM algorithm is developed that enables sparse variable selection while ensuring accurate parameter estimation.

The resulting estimator is shown to have the oracle property, and a simulation study is performed and confirms the strong finite-sample performance of the proposed method. Finally we apply the proposed approach to a set of real data on medical non-adherence arising from the Sequenced Treatment Alternatives to Relieve Depression Study and identify some new risk factors.

Regression Analysis of Multiplicative Hazards Model with Time-Dependent Coefficient for Sparse Longitudinal Covariates

⋆ *Zhuowei Sun, Hongyuan Cao*

Dalian Medical University, Florida State University

This work interests include survival analysis, longitudinal data modeling, and dynamic risk prediction. It focuses on the multiplicative hazards model with time-varying coefficients and intermittently observed longitudinal covariates. To address the bias inherent in ad hoc approaches such as the last value carried forward, the authors propose a kernel-weighted estimation method that yields unbiased estimation of the nonparametric coefficient function. Asymptotic normality is established at any fixed time point, and simultaneous confidence bands are constructed to assess the overall magnitude of variation. Theoretical findings are supported by simulation studies, and the proposed methodology is illustrated using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

S025: AI and Big Data Analysis

Global Inference for Distribution Functions in Large-Scale Data: A Bootstrap Approach with Distributed Computation

Li Cai

Zhejiang Gongshang University

The cumulative distribution function (CDF) is a fundamental statistical tool for characterizing random variables. Determining whether data follow a normal or heavy-tailed distribution is critical for guiding analytical choices and ensuring reliable subsequent inferences. In this paper, we focus on making global inferences for CDF across its entire support in large-scale data settings by constructing simultaneous confidence bands (SCBs). Existing methods are designed for single-machine computation

and rely on asymptotic limiting distributions, which often lead to conservative inference in finite samples due to the finite sample distortion of the limiting approximation. To address these limitations, we propose novel distributed and pseudo-distributed bootstrap methods for global inference on the large-scale data based on K-S statistics. Specifically, a distributed smooth kernel estimator is employed to estimate the distribution function of large-scale data, and the bootstrap method is used to avoid asymptotic approximations, thereby improving finite-sample accuracy. The proposed method also accommodates the empirical distribution function as a special case when the bandwidth is set to zero. The pseudo-distributed bootstrap is particularly advantageous for extremely large datasets, as it directly resamples subset-level statistics. The distributed bootstrap method permits the number of partitioned subsets to be either finite (small) or asymptotically large (tending to infinity), whereas the pseudo-distributed bootstrap method requires the number of subsets to be sufficiently large (also tending to infinity). Simulation studies confirm the theoretical findings and demonstrate the superior performance of our methods. As an illustration, we apply the proposed method to two real-world datasets: the CardioGoodFitness retail data and a dataset on developer productivity with AI and burnout.

Adaptive Allocation Method for Large Portfolio

Haitao Tian, Shouri Hu, ◆ *Lei Huang*

Southwest Jiaotong University

The mean-variance framework is a cornerstone of portfolio allocation, yet its high dimensional application is hindered by the challenge of accurate covariance matrix estimation. While sparse regression methods like the maximum-Sharpe-ratio estimated and sparse regression method (MAXSER) o er a promising framework, their reliance on the assumption of a constant covariance matrix ignores crucial time-varying market dynamic s. To overcome this limitation, we propose a novel method that integrates the MAXSER framework with a Dynamic Conditional Correlation GARCH model (DCC-GARCH). Our approach utilizes a time-varying covariance matrix to estimate the response variable, there by more accurately capturing evolving market conditions. Theoretical analysis con rms the enhanced risk-return pro le achieved through this integration. Comprehensive simulations and empirical studies on financial datasets demonstrate that the proposed DG-MAXSER outperforms both the classical mean-variance model and the standard MAXSER method, achieving a superior Sharpe ratio and more effective risk control. This method provides a robust and superior alternative for optimizing high-dimensional portfolios.

Copula-Enhanced Vision Transformer for High Myopia Diagnosis Through OU UWF Fundus Images

Chong Zhong, Yunhao Liu, Yang Li, Xiang Fu, Jin Yang, Danjuan Yang, Meiyan Li, Xingtao Zhou, Bo Fu, ◆ *Catherine Liu*

The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, Fudan University, Fudan University, The Hong Kong Polytechnic University, Eye & ENT Hospital, Fudan University, Eye & ENT Hospital, Fudan University, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

The advancement of AI-assisted myopia screening necessitates the joint diagnosis of pair-eye binary high myopia (HM) status and the prediction of continuous axial length (AL). This clinical requirement introduces a complex mixed-type multi-response learning task with multi-domain image covari- ates, resulting in two key challenges: i) capture the inter-ocular asymmetry within a cutting-edge foundation model; ii) model and estimate the condi- tional dependence structure among multivariate mixed-type responses given image covariates. We address the challenges by: i) imposing residual adapters to the Vision Tranformer foundation model to model the OU proximity and heteroskedasticity; ii) developing a copula loss that is available in PyTorch based on a latent variable expression, and proposing a computationally ef- ficient Fast Monte Carlo Expectation Maximization (fMCEM) algorithm to estimate copula parameters. We observe a new overfitting issue named higher covariance phenomenon in multitask learning, which has a substantial disturbance to estimation of copula parameters. The application to our annotated OU ultra-widefiled fundus image dataset and simulation studies demonstrate the superiority of our method in predictive capabilities.

PB-DiffHiC: a Statistically Principled Method for Detecting Differential Chromatin Interactions Using raw Pseudo-Bulk Hi-C Data

◆ *Yan Zhou, Yaohua Hu, Liuting Tan, Jiadi Zhu, Yutong Fei, Ming Gu, Dechao Tian*

Shenzhen University, Shenzhen University, Shenzhen University, Shenzhen University, Shenzhen University, Sun Yat-sen University (Shenzhen Campus), Sun Yat-sen University (Shenzhen Campus)

Single-cell Hi-C data provide unprecedented opportunities for analyzing differential chromatin interactions, essential for understanding genome structure-function relationships across various biological conditions. However, differential chromatin interaction analysis requires Hi-C data at high resolutions (e.g., 10 Kb), and due to the sparsity of scHi-C data, existing methods typically rely on single cell imputation or conventional bulk approaches, which can comprise the reliability of differential interaction detection. Here, we present PB-DiffHiC, a new optimized parametric statistical frame work that directly analyzes

the raw pseudo-bulk Hi-C data at 10 Kb resolution between conditions. PB-DiffHiC incorporates Gaussian convolution, stability of short-range interactions, and Poisson distribution to enable joint normalization and detection of significant differential chromatin interactions.Benchmark using cell-type-specific chromatin loops shows that PB-DiffHiC achieves higher precision than alternative methods. Applying PB-DiffHiC to pseudo-bulk and matched bulk Hi-C data demonstrate stronger concordance with identified differential interactions, reinforcing reliability. In a case study, PB-DiffHiC successfully identifies Kcnq5-associated differential interactions, closely matching SnapHiC-D results despite not relying on single-cell imputation. Therefore, PB-DiffHiC is a statistically sound method for enhanced comparative analysis of chromatin interactions using raw pseudo-bulk Hi-C data at 10 Kb resolution. The source code of PB-DiffHiC is publicly available at https://github.com/Tian-Dechao/PB-DiffHiC.

S030: Bootstrap and Model Averaging Methods for Robust Statistical Learning

Bootstrap Model Averaging

◆ *Minghui Song, Guohua Zou, Alan T.K. Wan*

Beijing Wuzi University, Capital Normal University, City University of Hong Kong

Model averaging has garnered significant attention in recent years for its ability to combine information from multiple models. A critical challenge in frequentist model averaging is determining the appropriate weight vector. The bootstrap method, well-known for its desirable properties, offers a promising solution. In this paper, we propose a bootstrap model averaging approach that selects weights by minimizing a bootstrap-based criterion. Notably, our weight selection criterion can also be interpreted as bootstrap aggregating. When all candidate models are misspecified, we show that the resulting estimator is asymptotically optimal, achieving the minimum possible squared error loss. Furthermore, we establish the convergence rate of the bootstrap weights toward the theoretically optimal weights. In scenarios where correct candidate models exist within a nested set and the number of covariates is fixed, we derive the limiting distribution of our proposed model averaging estimator. Through simulation studies and empirical applications, we demonstrate that our proposed method often outperforms other commonly used model selection and model averaging techniques, and other bootstrap-based variants.

Model Screening in Model Averaging: How and When?

◆ *Guangyuan Cui, Alan Wan, Xinyu Zhang*

City University of Hong Kong Shenzhen Research Institute, City University of Hong Kong, Academy of Mathematics and Systems Science

This paper targets a long-standing blind spot in the literature of model averaging: how to propose a suitable model screening that benefits model averaging with solid theoretical guarantees. Earlier studies mostly used screening to trim the candidate set, while seldom proving whether the trimmed estimator rivals or even surpasses the benchmark based on the full model set. Building on the \textit{full asymptotic optimality} property, we develop a new, risk-minimization-oriented screening framework that selects, among all model subsets of the same size, the one that minimizes the corresponding model averaging risk. The resulting estimator theoretically dominates other competing screening estimators of the same cardinality and is shown to attain the full asymptotic optimality under mild conditions. A consistent plug-in version of the proposed screening is provided for practical use and a dynamic-programming algorithm is proposed for efficient implementation of both screening methods. Furthermore, we demonstrate when the appropriate use of model screening accelerates the convergence rate of the asymptotic loss optimality. Finally, we discuss the circumstances where it is suggested to utilize model screening before model averaging. Extensive simulation studies confirm the advantages of the proposed screening methods and other theoretical findings.

High-Resolution Feature Identification in High-Dimensional Clustering

Lyuou Zhang

Shanghai University of Finance and Economics

Interpretable clustering, which involves simultaneous identifying heterogeneous subpopulations and the informative features that define the subpopulations, is a critical yet challenging task across various fields, including omics studies, clinical research, and policy evaluation. Existing methods typically either focus narrowly on global feature heterogeneity or treat feature identification and clustering as separate tasks, failing to account for their interaction. To address these limitations, we propose a novel unsupervised learning approach, PAirwise REciprocal fuSE (PARSE), which concurrently pinpoints cluster-specific informative features and conducts high-dimensional clustering effectively.

Sign Hacking in The Age of Big Data

◆ *Jie Hu, Yuhong Yang*

Tsinghua University, Tsinghua University

In linear regression, coefficient signs conveys the direction of covariate effects and are central to empirical interpretation. In high-dimensional settings, however, the abundance of candidate

covariates introduces substantial model selection uncertainty. We study the deliberate manipulation of coefficient signs through the inclusion of a carefully chosen auxiliary variable, a practice we term SHAVE (\textit{Sign Hacking with Auxiliary Variable Exploration}). We show that, conditional on the outcome and variables of interest, there exists a set of auxiliary-variable realizations with positive Lebesgue measure that induces sign reversals upon inclusion. Moreover, with high probability, such variables can be found when many auxiliary candidates are available, leading simultaneously to revered signs, inflated t- and F-statistics. Simulation studies and an empirical application corroborate these theoretical findings. We further propose detection strategies for SHAVE when independent datasets are available, as SHAVE has important implications for reproducibility, p-hacking, and research integrity.

S041: Advances in Statistical Inference for Dependent and High-Dimensional Data

Precise Asymptotics and Refined Regret of Variance-Aware UCB

⬩ *Yingying Fan, Yuxuan Han, Jinchi Lv, Xiaocong Xu, Zhengyuan Zhou*

University of Southern California

Abstract: In this paper, we study the behavior of the Upper Confidence Bound-Variance (UCB-V) algorithm for the Multi-Armed Bandit (MAB) problems, a variant of the canonical Upper Confidence Bound (UCB) algorithm that incorporates variance estimates into its decision-making process. More precisely, we provide an asymptotic characterization of the arm-pulling rates for UCB-V, extending recent results for the canonical UCB in Kalvit and Zeevi (2021) and Khamaru and Zhang (2024). In an interesting contrast to the canonical UCB, our analysis reveals that the behavior of UCB-V can exhibit instability, meaning that the arm-pulling rates may not always be asymptotically deterministic. Besides the asymptotic characterization, we also provide non-asymptotic bounds for the arm-pulling rates in the high probability regime, offering insights into the regret analysis. As an application of this high probability result, we establish that UCB-V can achieve a more refined regret bound, previously unknown even for more complicated and advanced variance-aware online decision-making algorithms.

Fast Community Detection of Discrete-time Temporal Networks in Mixture Dynamic Stochastic Block Models

Yang Liu, ⬩ *Binghui Liu*

Northeast Normal University, Northeast Normal University

Detecting communities of discrete-time temporal networks in dynamic stochastic block models has garnered significant attention. These models combine static stochastic block models with Markov chains of community label variables, which allows community variables to evolve over time. However, all existing studies on such models assume that the community labels of all nodes independently follow the identical Markov chain. This assumption is rather restrictive, particularly for temporal networks where nodes display distinct heterogeneity in their dynamic community structures. Hence, we establish the mixture dynamic stochastic block model, stating that the time series of community labels follow a mixture Markov chain with multiple components. Fitting this model is a challenging issue. To tackle this issue, we propose a mixture dynamic profile-pseudo-likelihood framework based on a decoupling strategy and a thoughtfully designed surrogate of the likelihood function. The proposed framework offers advantages in terms of computational efficiency, community recovery and prediction accuracy, as well as mixture component recovery accuracy, while also boasting solid theoretical guarantees for both the convergence of the framework and the consistency of the resulting parameter estimators. Furthermore, we extend it to handle discrete-time dynamic networks with degree heterogeneity. Finally, through extensive numerical results, we illustrate the feasibility and merits of the proposed framework.

Statistical Inference for Power Autoregressive Conditional Duration Models with Stable Innovations

⬩ *Yuxin Tao, Dong Li*

Southern University of Science and Technology, Tsinghua University

Financial data often exhibit irregular intervals in terms of transaction time, and the study of transaction timing has become increasingly important in empirical finance. However, the durations between financial events sometimes display heavy-tailed characteristics, which current duration models fail to capture. In this paper, we propose a first-order power autoregressive conditional duration model with positive alpha-stable innovations (sPACD), which effectively addresses the excess kurtosis in the durations. Further, the power form of the model structure mitigates the issue of overpredicting short durations in the classical ACD model. We study the properties of maximum likelihood estimation (MLE) within a unified framework of stationary and explosive cases, and discuss the estimation of the asymptotic covariance matrix. Strict stationarity test statistics and a modified Kolmogorov-type test statistic are established for stationarity testing and diagnostic checking in both stationary and explosive scenarios. Monte Carlo simulation studies demonstrate the good performance of the MLE and test statistics in finite samples. An empirical example is analyzed to illustrate the usefulness of sPACD models.

S045: Frontiers in Biostatistics: Robust Inference and Machine Learning

Heterogeneous Latent Variable Modeling of Multivariate Outcomes

Yi Liu, John Hanfelt, ⋆ *Limin Peng*

Emory University, Emory University, Emory University

In biomedical studies, multiple outcomes are often measured on the same individuals, providing complementary views of health status. Jointly analyzing these outcomes and understanding their heterogeneity can provide more robust assessment of disease progression and help guide customized treatment decisions. In this work, we propose to model the heterogeneity in multivariate outcomes by formulating a latent individual frailty and linking it to observed covariates. We develop an efficient estimation procedure based on the conditional score principle. Our modeling and estimation framework can flexibly accommodate continuous, binary, and categorical outcomes. The proposed method demonstrates satisfactory finite-sample performance in numerical studies.

Bilevel Gene Network Estimation with a Deep Neural Network Under Hierarchical Structured Sparsity

Jiayi Fan, Jingyuan Yang, Shuangge Ma, ⋆ *Mengyun Wu*

Shanghai University of Finance and Economics, Shanghai University of Finance and Economics, Yale School of Public Health, Shanghai University of Finance and Economics

Accurate gene network estimation serves as the cornerstone for understanding complex biological systems. Modern applications demand methods that can simultaneously capture nonlinear dependencies among genes and reconstruct the inherent hierarchical organization within biological networks, where higher-level entities such as functional pathways coordinate lower-level components like gene clusters. Traditional Gaussian graphical models are fundamentally limited in these aspects due to their linear assumptions and flat network representations. We propose a deep learning framework for bi-level network inference. The core innovation lies in its hierarchical selection layers that enforce structural consistency between high-level coordinating groups and their constituent low-level connections via adaptive sparsity constraints. This architecture is integrated with compositional deep models that learn cross-level association patterns through constrained nonlinear transformations, thereby explicitly preserving hierarchical dependencies while overcoming the representational limitations of linear methods. Crucially, we establish formal theoretical guarantees for the consistent recovery of both high-level connections and their internal low-level structures under general statistical regimes. Extensive validation demonstrates the effectiveness of our framework across synthetic and real-world scenarios, achieving superior F1 scores compared to competitive methods and particularly benefiting complex systems analysis through its interpretable bi-level structure discovery.

On Multiple Robustness of Proximal Dynamic Treatment Regimes

⋆ *Yuanshan Gao, Yang Bai, Yifan Cui*

Zhejiang University, National University of Singapore, Zhejiang University

Dynamic treatment regimes are sequential decision rules that adapt treatment according to individual time-varying characteristics and outcomes to achieve optimal effects, with applications in precision medicine, personalized recommendations, and dynamic marketing. Estimating optimal dynamic treatment regimes via sequential randomized trials might face costly and ethical hurdles, often necessitating the use of historical observational data. In this work, we utilize proximal causal inference framework for learning optimal dynamic treatment regimes when the unconfoundedness assumption fails.

Our contributions are four-fold: (i) we propose three nonparametric identification methods for optimal dynamic treatment regimes; (ii) we establish the semiparametric efficiency bound for the value function of a given regime; (iii) we propose a (K+1)-robust method for learning optimal dynamic treatment regimes, where K is the number of stages; (iv) as a by-product for marginal structural models, we establish identification and estimation of counterfactual means under a static regime.

Numerical experiments validate the efficiency and multiple robustness of our proposed methods.

A Machine Learning-Based Variance Reduction Approach of Average Treatment Effect Estimation in Observational Studies

Hao Mei, Siqi Liu, ⋆ *Zhibo Cai*

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

The inverse probability of treatment weighted (IPTW) estimator is widely used for estimating average treatment effects (ATEs) in observational studies. Although the IPTW method is advantageous in its ability to eliminate confouding, increased variance due to variability in the estimation of the propensity score is often a major concern. In this study, we propose a machine learning predictor-adjusted IPTW (ML-IPTW) estimator, where the weighted least squares estimator of the ATE is derived from a linear model that adjusts for machine learning predictors of the outcome based on covariates. Advanced from the traditional IPTW method, machine learning techniques are employed to reduce variance, incorporating covariate high-dimensionality

and non-linearity. Theoretically, we establish the consistency and asymptotic normality of the proposed estimator. We also derive the closed-form asymptotic variance and provide a consistent plug-in estimator. We demonstrate that the ML-IPTW estimator remains robust even when the machine learning step yields poor predictions, performing asymptotically no worse than the standard IPTW estimator. Empirically, our simulations and data analysis show that the proposed method achieves substantial variance reduction in the unbiased estimation of ATE.

S129: Statistical Foundations for Preference Learning, Representation, and Heterogeneity in the Era of AI

When Less Is More: Binary Feedback Can Outperform Ordinal Comparisons in Ranking Recovery

• *Shirong Xu, Jingnan Zhang, Junhui Wang*
Xiamen University, University of Science and Technology of China, The Chinese University of Hong Kong

Paired comparison data, where users evaluate items in pairs, play a central role in ranking and preference learning tasks. While ordinal comparison data intuitively offer richer information than binary comparisons, this paper challenges that conventional wisdom. We propose a general parametric framework for modeling ordinal paired comparisons without ties. The model adopts a generalized additive structure, featuring a link function that quantifies the preference difference between two items and a pattern function that governs the distribution over ordinal response levels. This framework encompasses classical binary comparison models as special cases, by treating binary responses as binarized versions of ordinal data. Within this framework, we show that binarizing ordinal data can significantly improve the accuracy of ranking recovery. Specifically, we prove that under the counting algorithm, the ranking error associated with binary comparisons exhibits a faster exponential convergence rate than that of ordinal data. Furthermore, we characterize a substantial performance gap between binary and ordinal data in terms of a signal-to-noise ratio (SNR) determined by the pattern function. We identify the pattern function that minimizes the SNR and maximizes the benefit of binarization. Extensive simulations and a real application on the MovieLens dataset further corroborate our theoretical findings.

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

Jiarui Zhang, Shuoxun Xu, • *Jiasheng Shi, Xinzhou Guo*
The Chinese University of Hong Kong, Shenzhen, The Hong Kong University of Science and Technology

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, and outperforms unsupervised benchmarks in predictive accuracy. 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.

Learning Coefficient Heterogeneity with Hierarchical Structure Information

• *Yan Zhong, Qiao Zheng, Lei Ma, Jianhua Huang*
East China Normal University, Shanghai University of International Business and Economics, East China Normal University, The Chinese University of Hong Kong, Shenzhen

Identifying coefficient heterogeneity across observed units from multiple latent subpopulations is a key challenge in many modern regression problems. Hierarchical relations among units provide valuable structural information that can substantially guide this task, yet existing methods rarely exploit this kind of information. We propose Hier-CCR, a regression framework that utilizes known hierarchical information to jointly estimate unit-specific coefficients and identify clusters of units with shared coefficients. Hier-CCR introduces a hierarchical Lasso penalty informed by a binary tree to incorporate the hierarchy. We establish coefficient estimation consistency and cluster recovery consistency for Hier-CCR under a broad class of loss functions, where both the number of observed units and the number of latent subpopulations grow with the sample size. We further develop a communication-efficient algorithm for Hier-CCR in decentralized federated systems with hierarchical structure. Simulation studies and an application to cell-type-specific gene regulation illustrate that Hier-CCR can recover heterogeneous coefficients and latent clusters effectively.

A Statistical Framework for Alignment with Biased AI Feedback

• *Xintao Xia, Zhiqiu Xia, Linjun Zhang Zhang, Zhanrui Cai*
Zhejiang University, Rutgers University, The University of Hong Kong

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.

S054: Inference and Learning Under Heterogeneous Data

Online Inference Under Over-Parameterized Models with Hidden Confounders

Shuyan Chen, ⋆Xingdong Feng, Yeheng Ge, Tao Li, Mengyun Wu
University of Science and Technology, Shanghai University of Finance and Economics, The Hong Kong Polytechnic University, Shanghai University of Finance and Economics, Shanghai University of Finance and Economics

We explore the online estimation and inference of coefficients in the presence of hidden confounders leveraging the over-parameterized models. Unlike existing offline approaches that rely on factor and sparse models, our closed-form estimator simultaneously removes hidden-confounder bias and is immediately usable for streaming data. Furthermore, with the aid of random matrix theory, we study the phase transition phenomena for the variance of the coefficient estimator that occur when the number of observations transitions from being fewer than to exceeding the number of predictors. Importantly, we demonstrate that adding more covariates slightly affects the estimated variance of the estimator, alleviating concerns about variance explosion within the over-parameterized model. The effectiveness of our approach for individual coefficient inference and multiple testing is validated through simulations and practical applications to two real datasets.

Heterogeneous Quantile Treatment Effect Estimation for Longitudinal Data with High-Dimensional Confounding

Zhixin Qiu, Huichen Zhu, Wenjie Wang, ⋆Yanlin Tang
East China Normal University, The Chinese University of Hong Kong, Eli Lilly and Company, East China Normal University

Causal inference plays a fundamental role in various real-world applications. However, in the motivating non-small cell lung cancer (NSCLC) study, it is challenging to estimate the treatment effect of chemotherapy on circulating tumor DNA (ctDNA). First, the heterogeneous treatment effects vary across patient subgroups defined by baseline characteristics. Second, there exists a broad set of demographic, clinical and molecular variables act as potential confounders. Third, ctDNA trajectories over time show heavy-tailed non-Gaussian behavior. Finally, repeated measurements within subjects introduce unknown correlation. Combining convolution-smoothed quantile regression and orthogonal random forest, we propose a framework to estimate heterogeneous quantile treatment effects in the presence of high-dimensional confounding, which not only captures effect heterogeneity across covariates, but also behaves robustly to nuisance parameter estimation error. We establish the theoretical properties of the proposed estimator and demonstrate its finite-sample performance through comprehensive simulations. We illustrate its practical utility in the motivated NSCLC study.

A Transfer Learning Framework for Efficient Portfolio Selection

Hongmei Lin
Shanghai University of International Business and Economics

Portfolio selection plays a critical role in financial decision-making by helping investors optimize returns, meet financial objectives, and manage risk. However, practical applications often face challenges such as limited data and high noise levels, particularly in emerging markets. This paper investigates the potential of transfer learning to address these issues in portfolio selection. We propose a general framework that incorporates information from auxiliary sources through a proposed transferable source detection algorithm. We further provide a rigorous theoretical analysis, establishing l_1-norm estimation error bounds and consistency guarantees for the proposed methods. Empirical results on both simulated and real-world datasets demonstrate that our transfer learning approach substantially improves portfolio performance, yielding higher Sharpe ratios and more efficient asset allocation, especially when more source data are available or a logarithmic utility function is employed.

Federated R-Learner for Heterogeneous Treatment Effect Estimation

Runyan Xin, Yiqiao Jin, ⋆Nan Lin, Linying Zhang
Washington University in St. Louis, Washington University in St. Louis, Washington University in St. Louis, Washington University

in St. Louis

Estimating heterogeneous treatment effects from distributed data is essential for precision medicine but challenging due to privacy constraints. We propose a Federated R-learner (Fed-R) that combines local cross-fitting for nuisance estimation with communication-efficient aggregation, enabling CATE estimation without sharing patient-level data. The results show that Fed-R achieves performance comparable to centralized methods under favorable conditions. This provides a practical and transparent framework for federated causal inference in distributed health data.

S056: Innovative Approaches in Machine Learning and Health Data Analytic

Cognition-Guided Regional Amyloid Thresholds Detect Earlier Domain-Specific Decline in Cognitively Normal Adults

Chung Chang

Sun Yat-sen University, Taiwan

Global amyloid-β (Aβ) positivity is often defined by a single summary SUVR cut-off, which may overlook regional deposition and its distinct relationships with domain specific cognitive change in cognitively normal adults. The aim of this study is to develop and evaluate a cognition-guided framework for determining global and regional Aβ SUVR cut-offs and to assess their ability to detect longitudinal cognitive change. We propose a method to estimate the optimal Aβ threshold within a linear mixed-effects model, while adjusting regression coefficient estimates and correcting p-values.

An Interpretable Hybrid Framework: Integrating Probabilistic Generative Modeling with Weighted Logistic Regression for Imbalanced Biomedical Data

⬩*Charlotte Wang, Yu-Ting Huang*

Taiwan University, Taiwan University

Researchers in biomedicine and public health increasingly face the challenges of class imbalance. This phenomenon is particularly evident in critical applications such as the serotype identification of vector-borne diseases, risk assessment of industrial fires and explosions, and large-scale cancer screening, where the events of primary interest are inherently rare yet carry significant clinical and social weight. While various hybrid and algorithm-level methods have been proposed, their inherent complexity and lack of transparency often limit their adoption in biomedical research. As a result, researchers favor data-level approaches; however, these approaches often fail to capture underlying distributions in complex scenarios, resulting in inconsistent classification performance.

We propose a hybrid framework that integrates Bayesian generative AI with subject-weighted logistic regression. Specifically, we introduce the BNNVAE, a model incorporating Bayesian Neural Networks into a Variational Autoencoder to generate highly representative synthetic samples that balance the dataset. To refine the classification process, we employ a subject-weighted logistic regression, in which weights are dynamically assigned based on sample representativeness and proximity to decision boundaries.

Simulations and real-world applications demonstrate that the proposed BNNVAE framework significantly outperforms traditional methods, such as SMOTE, achieving superior stability across F1-score, G-mean, and balanced accuracy metrics.

By leveraging Bayesian neural networks, the model effectively captures complex data distributions even with limited training data—a critical advantage in practical biomedical applications. Notably, our framework bridges the gap between performance and transparency; it generates highly representative synthetic data while maintaining an interpretable classification structure through sample re-weighting in the logistic regression. These findings underscore the framework's practical utility and scalability for addressing imbalanced data challenges in biomedical research.

A Prediction Platform to Estimate the Phytohormone Synthesis Potential in Bacteria Using Genomic Data

Zheng-Xiang Ye, ⬩*Steven Hung-Hsi Wu*

Taiwan University, Taiwan University

Indole-3-acetic acid (IAA) is a crucial plant hormone for growth and development. Among bacteria capable of IAA synthesis, tryptophan-dependent pathways are the primary pathway for IAA production. Identifying IAA-producing bacteria is essential for understanding plant–microbe interactions and developing sustainable agricultural practices. With the advancements in sequencing technologies which enable the acquisition of genomic data in a short amount of time at a reduced cost. This study developed a probabilistic framework to predict the likelihood of IAA synthesis in bacteria using their genomes, which offers a cost-effective screening approach compared to traditional wet-lab methods.

The prediction framework incorporates four tryptophan-dependent pathways and consists of two models. The Binary Model provides a binary prediction of IAA synthesis, and the Logistic Model ranks bacteria based on their synthesis potential. The models were validated against a comprehensive dataset of more than 15,000 bacterial genomes, and the results suggest that only 1.7% of the bacteria have the potential to synthesize IAA. The most promising phyla are Actinobacteria, Proteobacteria, and Firmicutes. The Logistic Model shows that the top-ranked genera include Streptomyces, Paraburkholderia, Nonomuraea, Rhodococcus, and Nocardia. The model also suggests that the potential for IAA

synthesis is likely to be strain-specific, emphasizing the importance of analyzing individual strains within a genus.

The proposed probabilistic prediction framework is a cost-effective in-silico approach for identifying bacteria with a high potential for IAA synthesis. This serves as a pre-screening tool for resource-intensive downstream wet-lab validation, significantly reducing the time and effort required to identify potential plant growth-promoting bacteria. Furthermore, the model can be customized for other pathways of interest, and making it a versatile framework to explore various bacterial metabolic processes.

A Data-Informed Variational Clustering Framework for Noisy High-Dimensional Data

Wanping Chen

Fu Jen Catholic University

Clustering in high-dimensional settings with severe feature noise remains challenging, especially when only a small subset of dimensions is informative and the final number of clusters is not specified in advance. In such regimes, partition recovery, feature relevance learning, and structural adaptation are tightly coupled, and standard likelihood-based methods can become unstable or overly sensitive to noisy dimensions. We propose DIVI, a data-informed variational clustering framework that combines global feature gating with split-based adaptive structure growth. DIVI uses informative prior initialization to stabilize optimization, learns feature relevance in a differentiable manner, and expands model complexity only when local diagnostics indicate underfit. Beyond clustering performance, we also examine runtime scalability and parameter sensitivity in order to clarify the computational and practical behavior of the framework. Empirically, we find that DIVI performs competitively under severe feature noise, remains computationally feasible, and yields interpretable feature-gating behavior, while also exhibiting conservative growth and identifiable failure regimes in challenging settings. Overall, DIVI is best viewed as a practical variational clustering framework for noisy high-dimensional data rather than as a fully Bayesian generative solution.

S058: Innovative Statistical Design Methods for Modern Clinical Trials

Posterior Predictive (PoP) Design for Next-Generation Phase I Clinical Trials

Chenqi Fu, ⋆Shouhao Zhou, J Jack Lee

Pennsylvania State University, Pennsylvania State University, UT MD Anderson Cancer Center

We introduce the posterior predictive (PoP) design, a novel interval-based design that leverages advanced Bayesian predictive hypothesis testing techniques for dose finding in Phase I cancer trials. Our work moves beyond the existing model-assisted interval-based designs by achieving global optimality in dose transition. Theoretically, the global optimality ensures that the proposed design can consistently select the true maximum tolerated dose (MTD) at an optimal square-root n convergence rate. Through extensive simulation studies, we demonstrate that the PoP design yields substantial improvement in operating characteristics to identify MTD, thereby presenting a valuable upgrade to the popular interval-based designs in practice.

A Pharmacokinetics-Informed Bayesian Design for Dose Optimization in Multi-Regional Clinical Trials

Xin Chen,Wenyun Yang, ⋆Fangrong Yan

China Pharmaceutical University, China Pharmaceutical University, China Pharmaceutical University

Multi-regional clinical trials (MRCTs) play a critical role in global drug development but face challenges in identifying region-specific optimal biological doses (OBDs) due to potential regional heterogeneity and limited sample sizes per region. We propose a Bayesian hierarchical model (BHM)-based design that incorporates pharmacokinetic (PK) data to guide borrowing strength across regions. The design evaluates two or more pre-specified doses in multiple regions, jointly modeling efficacy and toxicity via a Gumbel model. PK similarity across regions, assessed through a likelihood ratio test comparing nested PK models, informs the prior distribution of the shrinkage parameters of BHM. A utility function integrating efficacy-toxicity trade-offs supports OBD selection for both regional and overall populations. Simulation studies across homogeneous and heterogeneous scenarios demonstrate that the PK-informed BHM improves regional OBD selection accuracy compared to the independent analysis. In homogeneous settings, it achieves comparable performance to standard BHM while enhancing robustness in heterogeneous cases by reducing erroneous borrowing. This approach offers a practical framework for addressing regional inconsistency while maintaining statistical efficiency, supporting more reliable dose optimization for global drug development.

Metaheuristics for Finding Optimal Designs for Dose Response Trials

Weng-Kee Wong

University of California at Los Angeles

Metaheuristics is increasingly used across disciplines to tackle challenging optimization problems.

After a brief overview of nature-inspired metaheuristics, we discuss its recent applications to find various types of optimal

designs for clinical trials , including Locally Optimal Restricted Designs (LORDs) for phase I/II dose-finding studies that focus on both efficacy and toxicity outcomes. As an illustrative application, we find various LORDs for a 4-parameter continuation-ratio model and multi-stage designs for early phase trials. We also demonstrate the flexibility and usefulness of metaheuristics to tackle other optimization problems in the biomedical sciences.

Learning from Literature: Integrating LLMs and Bayesian Hierarchical Modeling for Next-Generation Clinical Trial Design

Wei Wei

Yale University

Designing modern oncology trials requires synthesizing evidence from prior studies to inform hypothesis generation and sample size determination. Trial designs based on incomplete or imprecise summaries can lead to misspecified hypotheses and underpowered studies, resulting in false positive or negative conclusions. To address this challenge, we developed LEAD-ONC (Literature to Evidence for Analytics and Design in Oncology), an AI-assisted framework that transforms published clinical trial reports into quantitative, design-relevant evidence. Given expert-curated trial publications that meet prespecified eligibility criteria, LEAD-ONC uses large language models to extract baseline characteristics and reconstruct individual patient data from Kaplan-Meier curves, followed by Bayesian hierarchical modeling to generate predictive survival distributions for a prespecified target trial population. We demonstrate the framework using five phase III trials in first-line non-small-cell lung cancer evaluating PD-1 or PD-L1 inhibitors with or without CTLA-4 blockade. Clustering based on baseline characteristics identified three clinically interpretable populations defined by histology. For a prospective randomized trial in the mixed-histology population comparing mono versus dual immune checkpoint inhibition, LEAD-ONC projected a modest median overall survival difference of 2.8 months (95 percent credible interval -2.0 to 7.6) and an estimated probability of at least a 3-month benefit of approximately 0.45. As LEAD-ONC remains under active development, these results are intended as preliminary demonstrations of the frameworks potential to support evidence-driven oncology trial design rather than definitive clinical conclusions.

S059: Integrative Statistical Methods for Genetic Association and Multi-Omics Studies

Modeling and Predicting Single-Cell Multi-Gene Perturbation Responses

Gefei Wang

Yale University

Understanding cellular responses to genetic perturbations is essential for understanding gene regulation and phenotype formation. While high-throughput single-cell RNA-sequencing has facilitated detailed profiling of heterogeneous transcriptional responses to perturbations at the single-cell level, there remains a pressing need for computational models that can decode the mechanisms driving these responses and accurately predict outcomes to prioritize target genes for experimental design. Here, we present scLAMBDA, a deep generative learning framework designed to model and predict single-cell transcriptional responses to genetic perturbations, including single-gene and combinatorial multi-gene perturbations. By leveraging gene embeddings derived from large language models, scLAMBDA effectively integrates prior biological knowledge and disentangles basal cell states from perturbation-specific salient representations. Through comprehensive evaluations on multiple single-cell CRISPR Perturb-seq datasets, scLAMBDA demonstrated robust generalization to unseen target genes and perturbations, and its predictions captured both average expression changes and the heterogeneity of single-cell responses. In addition, scLAMBDA can leverage information across datasets to generalize perturbation effects across multiple cell lines, extending its applicability beyond single-cell-line settings. We further support scLAMBDA's predictions using multiple orthogonal biological validations. Furthermore, its predictions enable diverse downstream analyses, including the identification of differentially expressed genes and the exploration of genetic interactions, demonstrating its utility and versatility.

Novel Multi-Context Transcriptome-Wide Association Framework Using Penalized Tensor Regression

*Ruilong Chen, Qiang Liu, Michael Epstein, *Jingjing Yang*

Emory University, Emory University, Emory University, Emory University

Existing single nucleus transcriptome-wide association study (snTWAS) only study one cell type at a time. To leverage reference gene expression profile across different cell types, we propose a novel snTWAS method based on penalized tensor regression that can jointly model gene expression of multiple cell types. For a target gene, its expressions in multiple cell types are arranged as a two-dimension tensor response, and its nearby cis-SNP genotype data are taken as predictors. To resolve the issue of more predictors than sample size in the tensor regression model, we develop a novel computation algorithm based on low-rank Tucker decomposition of the response tensor, with Elastic-Net regularization imposed on the factor matrices to induce sparsity and stabilize estimation. We applied to a snRNA-seq dataset of 436 prefrontal cortex samples comprised of six major brain cell types to perform snTWAS of Alzheimer's disease dementia (ADD). Our framework obtained valid gene expression prediction

models for 5x more genes (~14K) than the existing PrediXcan method. Our framework further detected 5x more significant risk genes than PrediXcan, including cell-type specific ones and the ones shared by multiple cell types. These TWAS risk genes were found enriched in immune activation, endocytosis, and lipid-processing pathways, and connected to known AD risk genes such as APOE, BIN1, PICALM, and PTK2B by protein-protein interaction analyses. Our novel tensor-based snTWAS method can model shared genetic architecture within transcriptomics data of multi-cell-types, while providing TWAS results for each cell type with superior performance.

Multi-resolution Clustering with Cross-Algorithm Reconciliation in Genomic Data Analysis

Xiaoquan Wen, Ali Turfah

University of Michigan, University of Michigan

Cluster analysis is a widely used unsupervised learning technique in data science, with critical applications in genomics such as inferring genetic population structures and annotating cell types from single-cell RNA-seq data. However, most existing clustering methods focus on identifying a single optimal partition while overlooking intrinsic relationships among the inferred clusters. Moreover, clustering results produced by different algorithms often appear inconsistent, and there is a lack of principled approaches to extract shared, scientifically meaningful patterns across diverse clustering outputs. In this presentation, we introduce a computational framework that enables systematic exploration of multi-resolution clustering structures in scientific data, starting from an initial configuration generated by any available clustering algorithm. The proposed algorithm provides a unified and principled approach for uncovering complex, nested latent structures and reconciling discrepancies among clustering results. Through comprehensive simulations and applications to large-scale, high-dimensional genetic and single-cell transcriptomic datasets, we demonstrate the framework's ability to recover interpretable and reproducible clustering patterns. Furthermore, we show that our multi-resolution cluster analysis of complex genomic data yields valuable insights into patterns of human population migration and cell differentiation trajectories.

S075: Modern Statistical Methods for High-Dimensional and Structured Data

Testing for Change Intervals in High-Dimensional Time Series

Kin Wai Chan, Ting Tin Ma

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

Different components of high-dimensional time series may respond to structural changes at varying rates, leading to change points clustering around an interval rather than occurring precisely at isolated points. We study a test for change intervals, or more broadly, change regions. The discrepancy from the null hypothesis is measured by the asymptotic proportion of coordinates that experience a change point over an unspecified change region. The test is constructed by comparing the empirical distribution of estimated change points across coordinates to the null distribution. This form of testing is highly scalable for high-dimensional time series and is tuning-free. The proposed framework covers the traditional change point test as a special case and avoids long-run variance estimation and self-normalization employed in mainstream testing approaches. The test statistic is proven to be pivotal asymptotically, and Monte Carlo simulations confirm good size control and high power even in finite samples.

Spatial Varying Coefficient Model with Covariate Measurement Errors: Bias-Corrected Inference and Efficient Computation

Bohai Zhang

Beijing Normal-Hong Kong Baptist University

Spatial varying coefficient (SVC) model offers an effective way to characterize the heterogeneous effects of covariates on the response variable. However, existing SVC models assume accurately measured covariates, ignoring real-world measurement errors that may arise from many aspects. This paper develops a statistically and computationally efficient framework for the inference of SVC models based on profile model likelihood, with explicit consideration of covariate measurement error. Focusing on spatial clustered coefficient models, we formulate an SVC model that allows functional covariate measurement errors. Bias-corrected estimators are derived by leveraging techniques of high-dimensional measurement error models and unbiased estimating functions. A predictive SVC model is then proposed to facilitate computations for large datasets. Finally, the effectiveness of the proposed approaches is demonstrated through comprehensive simulation studies and an analysis of the Arctic sea ice dataset.

Towards Understanding Massive Activations in Attention Sink Mechanism

Yuanyuan Lin

The Chinese University of Hong Kong

Recent studies have revealed two intriguing phenomena in large language models: attention sinks and massive activations. However, the co-emergence and co-existence of these two phenomena remain poorly understood. Inthis work, we revisit the prevailing view that massive activations are the primary mechanism responsible for concentrating attention on sink tokens,

and provide a more nuanced interpretation of their relationship. Through both theoreticalanalysis and empirical evidence, we demonstrate that massive activations and attention sinks jointly act to prevent excessive token mixing in self-attention. Specifically, attention sinks suppress mixing among non-sink tokens, whereas massiveactivations suppress mixing between sink tokens and non-sink tokens. Furthermore, our theory provides a principled explanation of how KV-biases, gating mechanisms and normalization layers can remove massive activations whilelargely preserving attention sinks. We further conduct intervention analyses and find that removing the value vector of the sink token can recover attention sinks even when massive activations are entirely suppressed. Overall, thiswork provides a mechanistic perspective on how massive activations and attention sinks interact under normalization and self-attention layers, offering new insights into their functional roles in Transformer models.

Errors-In-Variables Gaussian Processes for Mixed-Input Regression

Penghui Fu, ⋆Sheng Jiang

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

Mixed-input regression is prevalent in real-world applications. While Gaussian processes (GPs) have long been a fundamental tool for nonparametric regression, specifying an appropriate mixed-input covariance kernel remains challenging and is less well understood than its continuous-input counterpart. We propose an errors-in-variables (EIV) framework that imputes each qualitative input as a latent continuous variable. This reformulates the mixed-input regression into a joint model comprising a continuous-input GP in the latent space and a generative mechanism (e.g., a generalized linear model) for the qualitative inputs. This approach enables flexible, well-established kernel choice while preserving interpretability. Our framework is particularly well-suited for applications where qualitative inputs serve as inexpensive, noisy proxies for expensive, high-accuracy measurements of underlying physical state. Our innovations include (1) a unified modeling framework that generalizes existing mixed-input GPs; (2) an efficient Metropolis-within-Gibbs sampler for joint posterior inference and latent data imputation, providing natural uncertainty quantification; and (3) an optional design modeling strategy to enhance latent data estimation accuracy. Our framework seamlessly integrates into downstream tasks, providing natural uncertainty quantification, which is typically absent in mixed-variable Bayesian optimization and computer model calibration.

S076: Modern Statistical Methods for High-Dimensional Data and Machine Learning Models

Policy Evaluation in Distributional Reinforcement Learning via Kernel Mean Embeddings

Mehrdad Mohammadi, Qi Zheng, ⋆Ruoqing Zhu

University of Illinois Urbana Champaign, University of Louisville, University of Illinois Urbana Champaign

We propose an (offline) multi-dimensional distributional reinforcement learning framework (KE-DRL) that leverages Hilbert space mappings to estimate the kernel mean embedding of the multi-dimensional value distribution under a proposed target policy. In our setting, the state-action variables are multi-dimensional and continuous. By mapping probability measures into a reproducing kernel Hilbert space via kernel mean embeddings, our method replaces Wasserstein metrics with an integral probability metric. This enables efficient estimation in multi-dimensional state-action spaces and reward settings, where direct computation of Wasserstein distances is computationally challenging. Theoretically, we establish contraction properties of the distributional Bellman operator under our proposed metric involving the Matern family of kernels and provide uniform convergence guarantees. Simulations and empirical results demonstrate robust off-policy evaluation and recovery of the kernel mean embedding under mild assumptions, namely, Lipschitz continuity and boundedness of the kernels, highlighting the potential of embedding-based approaches in complex real-world decision-making scenarios and risk evaluation.

On Reference-Regulated Multiperiod Mean-Variance Portfolio Optimization in High Dimensions

Yutao Deng, Jianjun Gao, ⋆Weichen Wang

The University of Hong Kong

The multiperiod mean-variance (MV) portfolio optimization serves as a vital expansion of Markowitz's static MV portfolio selection framework. Just like its static counterpart, the multiperiod MV portfolio remains susceptible to estima- tion errors. We propose a reference-regulated multiperiod mean-variance (RRMV) framework that penalizes deviations from a reference policy. Therefore, this new optimization successfully combines the advantages of dynamic strategies and reference portfolios. A key contribution of this paper is the characterization of the out-of-sample Sharpe ratio under high-dimensional asymptotics with estimation errors in both the mean vector and the covariance matrix. We show how the reference penalty and the investment horizon jointly affect the optimized portfolio performance, and how regularization operates differently from the single-period portfolio optimization. Extensive simulation and real data studies demonstrate that the proposed framework improves the stability and out-of-sample Sharpe ratios of

multiperiod policies significantly.

Multi-View Orthogonal Projection Regression for Multi-Omics Data

Zongrui Dai, Yvonne Huang, • Gen Li

University of Michigan, University of Michigan, University of Michigan

Multi-omics data have strong within- and between-modality correlations, which pose a significant challenge for variable selection in regression analysis. We propose a novel method, Multi-View Orthogonal Projection Regression (MVOPR), for effective and interpretable variable selection in multi-omics studies. MVOPR incorporates directional relationships among omics layers to construct an orthogonalized feature space that removes redundancy and correlation among predictors. This transformation enables penalized regression to operate on decorrelated components, improving both selection accuracy and model interpretability. Comprehensive simulations demonstrate the efficacy of the proposed method. We further applied MVOPR to microbiome-metabolome data in an asthma study and identified key biologically relevant features associated with asthma.

S079: New Advances in Infectious Disease Modeling

Improving Assessment of Vaccine Effectiveness by Coupling Test-Negative Design Studies with Survival Models

Shangchen Song, Matt Hitchings, • Yang Yang, Ira Longini

University of Georgia

The test-negative design (TND) has become a widely used observational study design for evaluating vaccine effectiveness, especially during the COVID-19 pandemic. Traditionally, TND has been viewed as a variant of the case-control study and largely limited to use with logistic regression models. In this paper, we first establish that TND can be framed as a special case of a cohort study, thereby opening the door to a wider range of analytical approaches. We then introduce the Prentice, Williams, and Peterson gap-time (PWP-GT) frailty model as a novel method for analyzing TND data, accounting for recurrent infections and time-dependent vaccination status. Through extensive simulation studies, we demonstrate that the proposed model outperforms conventional models commonly applied in TND-based vaccine effectiveness studies. Finally, we apply our method to data from the National COVID Cohort Collaborative, estimating the effectiveness of full and booster doses of Pfizer's COVID-19 vaccines against both initial infection and reinfection during the Omicron variant circulation period in a real-world setting.

Estimating Real-Word Population-Level Vaccination Effects

Ryan Jasper, Nicholas Seewald, • Jing Huang

University of Pennsylvania, University of Pennsylvania, University of Pennsylvania

Estimating the population-level impact of vaccination coverage is critical for public health decision-making but presents substantial methodological challenges. Unlike individual-level vaccine effectiveness, population-level effects involve continuous, time-varying exposures and are not directly identifiable from randomized trials. We propose a unified framework that combines nested trial emulation with matching-based design and mixed-effects modeling to estimate the causal effect of incremental increases in vaccination coverage using observational data.

Specifically, we construct a sequence of weekly nested studies in which counties are grouped into subclasses based on similarity in time-varying covariates using optimal nonbipartite matching. Within each subclass, counties differ in vaccination coverage, enabling estimation of the effect of a hypothetical increase in coverage while controlling for confounding. We then fit a generalized linear mixed model that allows the effect of vaccination coverage to vary flexibly as a function of baseline coverage.

We apply this framework to U.S. county-level COVID-19 data from February to September 2021 to estimate the effect of a 1% increase in vaccination coverage on subsequent mortality. Results indicate that a 1% increase in coverage is associated with a 0.62% to 1.67% reduction in COVID-19 mortality over a two-week period, with consistent effects across baseline coverage levels. Sensitivity analyses varying lag structure and outcome windows yield similar results.

This framework provides a principled approach for estimating population-level causal effects under realistic conditions and can be broadly applied to evaluate public health interventions where randomized trials are infeasible.

Statistical and Machine Learning Approaches for Infectious Disease Dynamics

Tim Tsang

University of Hong Kong

Timely characterization of infectious disease transmission is critical for public health decision-making. In this talk, I present statistical and machine learning frameworks developed to address key challenges in the epidemiology of influenza, COVID-19, and other respiratory pathogens. I first describe Bayesian household transmission models that reconstruct person-to-person spread from community cohort data, enabling estimation of infectiousness across the clinical spectrum, including asymptomatic and pre-symptomatic infections, and identification of viral shedding kinetics as biomarkers for transmission risk. I then present methods for

reconstructing antibody dynamics from serial serological data to estimate individual infection risk and population-level immunity. Finally, I introduce an adaptive-weight ensemble approach that integrates mechanistic, statistical, and machine learning models to forecast influenza activity under the irregular seasonality patterns that emerged following COVID-19 disruptions. Throughout, I highlight how these methods have been deployed in real time to support public health responses in Hong Kong.

A Multiscale Spatio-Temporal Data Monitoring Framework for Prospective Infectious Disease Surveillance

Han Li, Xinyu Bi, ⋆Kai Yang, Shishun Zhao

Changchun University, Changchun University, Changchun University of Technology, Jilin University

Spatio-temporal process monitoring has broad applications in disease surveillance, spatial epidemiology, and many other areas. Most existing spatio-temporal process monitoring methods are developed for analyzing observed data at a single geospatial scale (e.g., state, county, or census tract), and usually impose some parametric assumptions on spatio-temporal data structure. For instance, observed data are often assumed to follow a parametric distribution (e.g., normal or Poisson) and the mean and covariance of a related spatio-temporal model are assumed to have parametric forms, which are rarely valid and/or difficult to justify in many applications. In practice, spatio-temporal data (e.g., incidence rates of infectious diseases) are usually managed at multiple geospatial scales and have complicated structure. In this talk, we introduce a new multiscale spatio-temporal data monitoring method for prospective disease surveillance. At each observation time, the new method first detects anomalous regions at each spatial scale. To reduce computing time, a forward elimination procedure is suggested for this purpose. Then, an optimal spatial scale at the current observation time is determined using the detected anomalous regions and a proposed scale information criterion. Finally, a charting statistic is constructed based on the detected anomalous regions at the optimal scale. This method is fully nonparametric in the sense that no parametric assumptions are imposed on the structure of observed spatio-temporal data. It is also scalable and computationally efficient, making it a reliable and effective tool for online multiscale spatio-temporal data monitoring problems.

S107: Recent Advances in Spatial Omics and Single-cell Data Analysis

Graph Transformer Inference Links Ligand-Receptor Signaling to Target Genes in Special Niches

Xiao Xiao, Yihuan Wang, Hongyu Zhao, Grace Bushnell, ⋆ Zuoheng Wang

Yale University, Yale University, Yale University, University of Minnesota, Yale University

Ligand-receptor (LR) interactions are a fundamental mechanism in cell-cell communication (CCC). Deciphering revelent LR pairs that regulate gene expression of receiver cells in aberrant CCC provides mechanistic insights for therapeutic target discovery. Integrating single-cell resolution spatial transcriptomics (ST) and single-cell RNA sequencing (scRNA-seq) data enables the prediction of ligand-receptor-target links between interacting cells by capturing both gene expression and spatial context. We present GITIII-AG, a computational tool that leverages ST and scRNA-seq data to infer upstream LR interactions that regulate gene expression of receiver cells. GITIII-AG first learns the influence of sender cells on the gene expression of receiver cells using an interpretable graph transformer trained by encoding cell state and niche correlation. It then examines the correlation between CCC influence and the signaling strength of LR pairs. Applied to datasets across platforms, GITIII-AG identified known drug targets and pinpointed new LR signaling that regulate target gene expression in breast cancer.

Background-Correction of Spatial Proteomics Data Using Finite Mixture of Gaussian Models

⋆*Agus Salim, Dharmesh Bhuva, Malvika Kharbanda*

University of Melbourne, University of Queensland, Adelaide University

Phenocycler technology can perform in-situ measurement of up to 45 protein markers at single-cell level. This technology allows researchers to study cell-level protein-protein interaction and identification of niches such as tumour microenvironment. Phenocycler intensity data is often bi- or tri-modal due to the presence of background f fluorescence and non-specific bindings. The existence of these unwanted signals reduces the quality of the data, thus reducing the potential of drawing meaningful conclusions. To alleviate this issue, we developed a class of Finite mixture of convoluted Gaussian distributions with three components that allows estimation of the true signals from the data. In the Gaussian mixtures, the first component is assumed to contain only the background fluorescence, the second component is the convolution of background and intensity originating from non-specific bindings and finally the third component is the convolution of the intensity originating from background, non-specific bindings and the true signals. Closed-form formulae for the estimates of true signals are derived without requiring knowledge of the correlation between background, non-specific binding and true intensities. We benchmark our methods using datasets of non-small cell lung cancer patients and demonstrated

that our automatic algorithm achieves similar segmentation accuracy accuracies as the time-consuming expert-based judgement.

Multi-Scale Learning Spatially Interpretable Latent Feature Representations in Spatial Transcriptomics

• *Yuanhua Huang, Jun Wang, Ole Winther, Tianjie Wang, Lingyu Li*

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

Spatial transcriptomics quantifies gene expression within its spatial context, significantly advancing biomedical research. Understanding gene spatial expression and the organization of multicellular systems is vital for disease diagnosis and studying biological processes. However, existing models often struggle to integrate gene expression data with cellular spatial information effectively. In this talk, we will first introduce SpatialFormer, which combines self-supervised training with auxiliary spatial supervisions of neighbouring cells or transcripts. By pre-training on 62 Xenium slides, we demonstrate that SpatialFormer distills biological signals across various tasks, including single-cell batch correction, cell-type annotation, co-localization detection, and identifying gene pairs that are critical for the immune cell-cell interactions involved in the regulation of lung fibrosis. Later, we will further show some ongoing work with unified frameworks to link spatial representation, cross-modality prediction, and biological interpretation.

BayesRare: A Bayesian Mixture Model for Population-Level Rare Cell Type Detection in Multi-Subject Single-Cell RNA Sequencing Data

• *Yinqiao Yan, Hao Wu*

Beijing University of Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences; Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology

Rare cell types in single-cell RNA sequencing (scRNA-seq) data often encode essential biological signals, such as early disease markers or key immune regulators. With advancing technologies, large-scale scRNA-seq cohorts from multiple subjects now enable population-level analyses of the prevalence, heterogeneity, and disease associations of rare cell populations. However, existing methods for rare cell detection are typically limited to single datasets and cannot effectively leverage cross-subject information. To tackle this challenge, we present BayesRare, a hierarchical Bayesian framework for population-level rare cell discovery in multi-subject scRNA-seq data. The method augments a Bayesian mixture model with a rare cluster indicator, supporting joint cell-type clustering and rare-population identification. By explicitly characterizing the statistical properties of rare cell types, BayesRare integrates evidence across subjects, quantifies uncertainty via posterior probabilities, and enables inference of group-level differences (e.g., patients versus controls). Across synthetic and three real datasets, BayesRare achieves superior precision, reduces false positives, and uncovers biologically meaningful disease-specific rare subtypes. The R package of BayesRare is available at https://github.com/yinqiaoyan/BayesRare.

S112: Recent Advances of Time Series Analysis and Change-Point Detection

Robust and Efficient Inference for Nonstationary GARCH(1,1) Models with Heavy-Tailed and Asymmetric Innovations

• *Huan Gong, Dong Li, Yuxin Tao*

National University of Defense Technology, Tsinghua University, Southern University of Science and Technology

Asset return volatility often exhibits heavy tails, asymmetry, and persistent dynamics, posing substantial challenges for inference in GARCH(1,1) models. Classical quasi-maximum likelihood estimation typically relies on Gaussian distributional assumptions and is usually studied under strict stationarity, so it may become inefficient or theoretically inadequate when these conditions are violated. We develop a generalized non-Gaussian quasi-maximum likelihood estimation (GNGQMLE) framework for robust and efficient inference in GARCH(1,1) models with heavy-tailed and possibly asymmetric innovations. The proposed method provides a unified treatment of stationary and nonstationary volatility dynamics under fractional-moment conditions. We establish consistency and asymptotic normality for the full parameter vector in the stationary case and for the dynamic parameter vector in the nonstationary case, thereby clarifying the regime-dependent distinction between estimable and inestimable parameters. By adaptively selecting a suitable non-Gaussian likelihood, the GNGQMLE improves efficiency while maintaining robustness to tail heaviness and asymmetry. We also propose a score-based portmanteau diagnostic test that remains valid in both stationary and nonstationary settings. Monte Carlo experiments illustrate the finite-sample performance of the proposed procedure, and an empirical application to daily cryptocurrency returns demonstrates its practical usefulness.

Adaptive Dynamic Pricing Under Non-Stationary Environments

• *Feiyu Jiang, Zifeng Zhao*

Fudan University, University of Notre Dame

Dynamic pricing strategies in online marketplaces usually cope

with evolving consumer demand driven by behavioral shifts, seasonality, and external disruptions. This paper studies adaptive dynamic pricing under a general non-stationary environment where the demand model may change abruptly at unknown times or drift gradually over time, or both. We propose a unified contextual dynamic pricing framework that requires no prior knowledge of the extent of non-stationarity and automatically adapts to both piecewise-stationary and variation-budget conditions. Theoretically, we establish that the resulting pricing strategy achieves $\min\{\sqrt{dTS}, \Delta^{1/3} T^{2/3}\}$ dynamic revenue regret, where d is the dimension of contexts, S is the number of abrupt changes, Δ is the quadratic variation budget of the demand parameters, and T is the selling horizon. These guarantees are minimax-optimal up to logarithmic factors in both piecewise-stationary and variation-budget regimes. Empirical studies demonstrate that our algorithm adapts effectively to diverse non-stationary patterns.

Two-Way Matrix Autoregressive Model with Thresholds

Cheng Yu, ⋆Xinyu Zhang, Dong Li, Howell Tong

The University of Chicago, East China Normal University, Tsinghua University, London School of Economics and Political Science

Recently, matrix-valued time series data have attracted significant attention in the literature with the recognition of threshold nonlinearity representing a significant advance. However, given the fact that a matrix is a two-array structure, it is unfortunate, perhaps even surprising, for the threshold literature to focus on using the same threshold variable for the rows and the columns. In fact, evidence in economic, financial, environmental and other data shows the necessity and advantages of allowing the possibilities of different threshold variables for the rows and columns, with possibly different threshold parameters, hence the need for a Two-way Matrix AutoRegressive model with Thresholds (2-MART). Naturally, two threshold variables pose new and perhaps even fierce challenges, which might be the reason behind the adoption of only one threshold variable in the literature up to now. In this paper, we develop a comprehensive methodology for the 2-MART model, by overcoming various challenges. Compared with existing models in the literature, the new model can achieve greater dimension reduction, much better model fitting, more accurate predictions, and more plausible interpretations.

Ranking Individuals in Team Tournaments: With an Application to E-Sport Games

⋆*Xuanling Yang, Yuxin Tao, Zixuan Sheng, Dong Li*

Beijing University of Chemical Technology, Southern University of Science and Technology, Tsinghua University, Tsinghua University

In classical ranking scenarios, comparisons are typically made directly between the items of interest. However, sometimes the individuals to be ranked cannot be compared directly. This paper addresses the ranking problem when the entities of interest are organized into teams, and only team-level comparisons are observed. We propose a new model based on the Bradley-Terry-Luce framework for the case where the number of members in each team is fixed. The proposed model introduces the concepts of conditional team score and conditional individual score. The former refines the team score based on the specific composition, while the latter captures the interactions among individuals. Finally, we apply the proposed method to a real-world E-sports match data to ranking the players and analyze their mutual influences.

S117: Recent Developments on Applied Statistics

Model-Robust Process Optimization for Data with a Multi-Stratum Structure

Xiaohua Liu, ⋆Po Yang, Chang-Yun Lin

University of Manitoba, University of Manitoba, Chung Hsing University

Response optimization is a process of identifying the input variable settings that optimize the response. Multistratum design arises naturally in industrial experiments due to the inconvenient and impractical completely randomization. Accounting for the model uncertainty, we apply the Bayesian model averaging method and the predictive approach to investigate the optimization problem for data with a multi-stratum structure. We use bridge sampling to estimate marginal likelihoods and hence posterior model probabilities. The goal of the optimization is to identify factor settings that maximize the probability that the response falls within a given range. The method is illustrated with two examples.

Shannon' s Entropy Estimation Under Undersampling

Tsung-Jen Shen

Chung Hsing University

Shannon's entropy is a fundamental metric for assessing species diversity. It is widely used in ecology because it captures the complexity of community structure without bias toward rare or abundant species. However, accurately estimating Shannon's entropy becomes challenging when species inventories are incomplete, a common issue in ecological research. This study develops a novel estimator to improve the precision and reliability of Shannon's entropy estimates under such conditions. In a simulation study, the proposed estimator was evaluated against

eight established estimators using diverse surveyed ecological datasets as true populations, including Malaya butterflies, beetles, small mammals, corals, and insects, across varying sample sizes. For most scenarios, the results show that the proposed estimator consistently outperforms traditional methods, particularly in small- and moderate-sized samples, by delivering lower bias and greater stability. This enhanced accuracy makes the new estimator a reliable tool for biodiversity assessment, with applications extending beyond ecology to fields such as medical and environmental science. The study provides a refined framework for biodiversity analysis, addressing critical challenges in ecological data collection and entropy estimation.

Clustering High-Dimensional Data with Missing Values Using Restricted Multivariate Skew-t Mixtures of Common Factor Analyzers

Tsung-I Lin, Wan-Lun Wang, Luis M. Castro, Shi-Xiu Yu

Chung Hsing University, Cheng Kung University, Pontificia Universidad Católica de Chile, Chung Hsing University

Mixtures of common restricted skew-t factor analyzers (MCrstFA) have emerged as a parsimonious and practical approach for the model-based clustering of high-dimensional data with asymmetric features and outlying observations in one or more heterogeneous populations. However, missing data has become a ubiquitous problem that frequently adds complexity to the analyses for practitioners. Most existing tools are not adaptable to data with missing values, which is a significant reason for this complexity. This paper proposes an extension of the MCrstFA framework that allows the analyst to parsimoniously model data with missing values, skewness, heavy tails, and multimodality simultaneously. Under the missing at random mechanism for nonresponses, a computationally feasible expectation conditional maximization either algorithm is developed for computing maximum likelihood estimates of model parameters. The estimation procedure enables the automatic imputation of missing values and the prediction of unobserved factor scores. To assess the precision of the ML estimators, an information matrix-based approach is employed to approximate the asymptotic covariance matrix of the estimators. Numerical results obtained from the analysis of simulated and real datasets illustrate the effectiveness of the proposed methodology.

Efficient and Robust Block Designs for Order-of-Addition Experiments

Chang-Yun Lin

Chung Hsing University

Designs for Order-of-Addition (OofA) experiments have received growing attention due to their impact on responses based on the sequence of component addition. In certain cases, these experiments involve heterogeneous groups of units, which necessitates the use of blocking to manage variation effects. Despite this, the exploration of block OofA designs remains limited in the literature. As experiments become increasingly complex, addressing this gap is essential to ensure that the designs accurately reflect the effects of the addition sequence and effectively handle the associated variability.

Motivated by this, this paper seeks to address the gap by expanding the indicator function framework for block OofA designs. We propose the use of the word length pattern as a criterion for selecting robust block OofA designs. To improve search efficiency and reduce computational demands, we develop algorithms that employ orthogonal Latin squares for design construction and selection, minimizing the need for exhaustive searches.

Our analysis, supported by correlation plots, reveals that the algorithms effectively manage confounding and aliasing between effects. Additionally, simulation studies indicate that designs based on our proposed criterion and algorithms achieve power and type I error rates comparable to those of full block OofA designs. This approach offers a practical and efficient method for constructing block OofA designs and may provide valuable insights for future research and applications.

S122: Some New Methods of Robust Statistical Learning for Complex Data

Local Linear QR Estimator of Conditional Quantile for Strong-Mixing and Truncated Data

Mei Yao,Jiangfeng Wang,Lu Lin, Yu Xia

Hefei University of Technology,Zhejiang Gongshang University,Shandong University, Hangzhou Normal University

Let $\{(X_i, Y_i), i \ge 1\}$ be a $d+1$-dimensional stationary strong-mixing process. When Y_i are randomly left truncated, a local linear QR estimator for conditional quantiles is proposed based on the quantile regression method. The Bahadur-type expression and asymptotic normality property of the estimator are established, which extends the results of Yu and Jones (1998) from one-dimensional covariate, independent and complete data to multivariate covariates, strong-mixing and truncated data. Finally, simulation studies are conducted to show that our method is more robust than the kernel methods proposed by Lemdani et al. (2009) and Yao et al. (2018).

Robust Subgroup Analysis for Heterogeneous Censored Data

Zhaohui Xu, Daoji Li, Zemin Zheng

University of Science and Technology of China, California State University, Fullerton, University of Science and Technology of

China

Subgroup analysis is important in practice because real-world data typically come from heterogeneous populations, where meaningful patterns can differ substantially across subpopulations. Correctly identifying these subgroups can improve prediction accuracy, prevent biased or misleading conclusions, and support more effective, targeted decision-making. While most existing subgroup analysis methods are developed for complete data, in this paper we propose a novel and robust approach for censored data under heterogeneous accelerated failure time (AFT) models. Specifically, we combine inverse probability weighting, M-estimation, and concave pairwise fusion penalization to simultaneously identify subgroups and estimate covariate effects for heterogeneous censored data, without requiring prior knowledge of individual subgroup memberships. We further develop an efficient RISA-ADMM algorithm to implement the method and establish its convergence. Furthermore, we derive the theoretical properties of the proposed estimators under mild regularity conditions. Extensive simulations and an application to the German credit dataset demonstrate the robustness and effectiveness of our approach.

Stable Semi-Distance Covariance for Independence Testing: From Low to High Dimensions

Feng Zou

Zhongnan University of Economics and Law

This paper introduces a Stable Semi-Distance Covariance (SSDCov) index to test the independence between a random vector $\mathbf{x} \in \mathbb{R}^p (p \geq 1)$ and a categorical variable Y. Rooted in the characteristic function of symmetric α-stable distributions, the SSDCov index is defined as the weighted integral of the squared difference between the conditional characteristic function of $\mathbf{x}$ given Y and its marginal characteristic function. A paramount theoretical advantage of SSDCov is that it serves as a generalized metric family that encompasses the existing semi-distance covariance (SDCov) as an asymptotic special case (when the bandwidth parameter $\sigma \to +\infty$) while significantly relaxing stringent finite moment assumptions. A key property is that $\mathrm{SSDCov}(\mathbf{x},Y) \geq 0$, with equality holding if and only if $\mathbf{x}$ and Y are independent. We systematically establish the asymptotic distributions of the proposed test statistic under both the null and alternative hypotheses across two distinct paradigms: the classical low-dimensional regime ($n \to \infty$ with fixed p) and the challenging high-dimensional setting ($p \to \infty$ with fixed n). To optimize testing performance, we further propose a data-driven bandwidth selection method via the median heuristic and derive the corresponding theoretical guarantees. Simulations demonstrate the superiority and striking stability of the proposed test over existing methods, especially for heavy-tailed data. Finally, the practicality of our method is illustrated through a real data analysis.

Robust Reproducibility Method Based on Modal Regression

Jinzhao Yu, Hui Zhao, Daoji Li

Hubei University of Economics

Although the modal regression has been widely used in practice, its false discovery rate(FDR) control problem remains unexplored. To address this gap, we propose a robust reproducibility method that aims to control the FDR during the variable selection procedure. Our approach leverages the robustness of modal regression. In incorporating the kernel-based objective function, we employ the regularized MEM algorithm and propose a data-driven method for bandwidth selection. In terms of theoretical contributions, we analyzed the upper bounds of FDR control from both finite-sample and asymptotic perspectives, and also established the asymptotic properties of the statistical power for the proposed method. Both theoretical analysis and experimental results demonstrate that our method achieves FDR control, especially for skewed distributions.

S124: Statistical Analysis of Complex Data

Combining Pre-Trained Models Via Localized Model Averaging

Ziwen Gao

Tsinghua University

Many pre-trained models (PTMs) are available in modern applications. Because different PTMs are often trained on different datasets, their performances can vary substantially for different new tasks, and the ranking of the candidates may depend heavily on the input. Motivated by this, 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 PTMs 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.

Frechet Regression with Mondrian Forests: Finite-Sample Guarantees and Ensemble Benefits

Rui Qiu, Fang Yao, Zhou Yu

Peking University, Peking University, East China Normal University

Fr\'echet random forests extend the power of classical random forests to general metric spaces, offering promising advantages over traditional methods, especially in high-dimensional settings. While forests have empirically outperformed individual trees in Fr\'echet regression, the theoretical basis for this improvement remains largely unexplored. This paper fills this gap by establishing non-asymptotic upper bounds for the prediction risk of Fr\'echet Mondrian forests, complementing the existing literature that primarily focuses on asymptotic analysis. We demonstrate that, under suitable regularity conditions, Fr\'echet Mondrian forests attain convergence rates comparable to their Euclidean counterparts. Moreover, under higher-order smoothness assumptions and with a sufficient number of trees, Fr\'echet forests achieve faster convergence than individual Fr\'echet trees, thereby providing a rigorous theoretical justification for the benefit of ensembles in non-Euclidean regression problems. The effectiveness of the proposed method is further corroborated through simulation studies across diverse settings, including probability distributions, symmetric positive-definite matrices, and spherical data.

Bias-Corrected Weighted Least Squares Model Averaging in Linear Measurement Error Models With Right Censored Data

Guozhi Hu

Hefei Normal University

This paper develops an optimal model averaging approach for linear measurement error models, when the response is randomly right censored. Within this context, we propose a bias-corrected weighted least squares estimation for unknown regression parameter. A novel Mallows-type weight choice criterion that skillfully bypasses the unavailable true covariates is developed for allocating model weights. We provide two theoretical justifications for our proposal. First, we establish the asymptotic optimality property of the resulting model averaging estimator when all candidate models are misspecified. Second, we show our model averaging method estimates the model parameter with root-n rate when the true model is among the candidate models. Particularly, it is demonstrated that the optimality is still valid when a model screening strategy is conducted prior to model averaging. Simulation studies and a real data analysis highlight the superiority of the proposed method.

Supervised Functional Tensor Autoregression

Songhua Tan

University of Science and Technology of China

Tensor data has attracted significant attention in statistical research. A common challenge with such data is the coexistence of temporal dependence and functional continuity in some modes. To address this, we novelly introduce a supervised functional tensor autoregressive (SFTAR) model where the functional tensor SVD decomposition is applied to reduces the dimension of autoregressive coefficient. Thus, the singular vectors and singular functions act as feature extractors, capturing latent patterns from observations. We design an iterative least squares method. Theoretically, we establish the non-asymptotic contractive error bounds for the proposed estimator. Simulations and a sea surface temperature analysis illustrate the finite sample performance of the proposed method.

S127: Statistical and Computational Methods for Single-Cell, Spatial, and Multi-Omics Integration

Designing Spatial Transcriptomics Using Histology-Driven Expression Prediction

Jean Yee Hwa Yang

The University of Sydney

Spatial transcriptomics technologies have enabled the quantification of spatially resolved gene expression, providing new insights into the molecular basis of disease. While many methodological approaches focus on modeling these spatially resolved data, relatively less attention has been given to the design of subcellular spatial transcriptomics experiments. In this presentation, we will discuss two of the key design tasks: the gene panel design and region of interest (ROI) selection. Panel design involves selecting genes for spatial transcriptomic assays to ensure that the resulting data effectively address the research question(s) of interest. Here, we introduce panelScope, a panel characterization framework that evaluates panels generated by human experts, computational methods, and large language models (LLMs). We found that while computationally constructed gene panels performed competitively in capturing major cell types when compared to our in-house manually curated panel; refined manual curation offered distinct other advantages such as capturing minor cell types. For ROI selection, we apply our newly developed algorithm, GHIST+, which first predicts spatial gene expression from cost-effective haematoxylin-and-eosin (H&E) stained histology images and then uses these predicted spatial expression profiles to identify the most informative regions. These predicted expression also enable us to digitally explore patterns of cell-type colocalization and potential intercellular communication within tumor microenvironments to guide experimental design.

IntegrateRigor: Statistical Framework for Rigorous Integra-

tion of Single-Cell and Spatial Transcriptomic Data Without Prior Annotations

Jingyi Jessica Li

Fred Hutchinson Cancer Center

IntegrateRigor is a statistical framework for evaluating and optimizing single-cell and spatial transcriptomics data integration. A central challenge in integration is balancing batch-effect removal against preservation of cell identity. Popular methods such as Harmony, Seurat, and scVI require ad hoc hyperparameter tuning and can under-integrate, leaving batch-driven separation, or over-integrate, collapsing distinct cell types and trajectories. Existing evaluation metrics either require cell-type annotations that are often unavailable, or, in the annotation-free case, have recently been shown to be unreliable. IntegrateRigor addresses this gap with two components. First, a per-gene batch stability score (BSS) identifies batch-stable genes and uses them as input to improve integration. Second, a dataset-level integration score combines a per-dimension batch alignment score (BAS) and a per-dimension cell identity score (CIS) to quantify, for each method and hyperparameter setting, how well the resulting embedding mixes batches while preserving cell identity. Neither component requires cell-type annotations. Together, these components enable researchers to select an integration strategy suited to their data, turning a trial-and-error process into a principled workflow that supports more reliable cell-type annotation, trajectory inference, and differential expression analysis.

FastCCC: A Permutation-Free Framework for Scalable, Robust, and Reference-Based Cell-Cell Communication Analysis in Single Cell Transcriptomics Studies

Siyu Hou, Wenjing Ma, •Xiang Zhou

Yale University, Yale University, Yale University

Detecting cell-cell communications (CCCs) in single-cell transcriptomics studies is fundamental for understanding the function of multicellular organisms. Here, we introduce FastCCC, a permutation-free framework that enables scalable, robust, and reference-based analysis for identifying critical CCCs and uncovering biological insights. FastCCC relies on fast Fourier transformation-based convolution to compute p-values analytically without permutations, introduces a modular algebraic operation framework to capture a broad spectrum of CCC patterns, and can leverage atlas-scale single cell references to enhance CCC analysis on user-collected datasets. To support routine reference-based CCC analysis, we constructed the first human CCC reference panel, encompassing 19 distinct tissue types, over 450 unique cell types, and approximately 16 million cells. We demonstrate the advantages of FastCCC across multiple datasets, most of which exceed the analytical capabilities of existing CCC methods. In real datasets, FastCCC reliably captures biologically meaningful CCCs, even in highly complex tissue environments, including differential interactions between endothelial and immune cells linked to COVID-19 severity, dynamic communications in thymic tissue during T-cell development, as well as distinct interactions in reference-based CCC analysis.

Computational Methods and Resources for Multi-Omics Understanding of Complex Human Systems

• *Yun Li, Jiawen Chen, Didong Li*

University of North Carolina, Chapel Hill, University of North Carolina, Chapel Hill, University of North Carolina, Chapel Hill

Spatial omics technologies revolutionize studies of tissue functions. However, existing methods fail to capture localized, sharp changes characteristic of critical events such as tumor development. We present StarTrail, a gradient based method that powerfully defines rapidly changing regions and detects “cliff genes”, genes exhibiting drastic expression changes at highly localized or disjoint boundaries. StarTrail, filling important gaps in current literature, enables deeper insights into tissue spatial architecture. We also introduce STimage-1K4M, a comprehensive dataset designed to bridge this gap by providing transcriptomic features for sub-tile images. STimage-1K4M contains 1,149 images and 4,293,195 pairs of sub-tile images and gene expressions. Spatial variable genes (SVGs), revealing critical information and about tissue architecture, cellular interactions, and disease-relevant microenvironments, are of keen interest in ST studies. We performed a comprehensive benchmarking study of 20 state-of-the-art SVG detection methods using >600 human slides from STimage-1K4M. We also constructed the first cross-tissue atlas of SVGs, enabling comparative analysis of spatial gene programs across cancer and normal tissues.

S130: Statistical Foundations of LLM Behavior, Alignment, and Automated Reasoning

Divergence Representations of LLM Safety Alignment

• *Qifan Song, Yue Xing, Rajdeep Haldar, Ziyi Wang, Guang Lin*

Purdue University, Michigan State University, Purdue University, Purdue University, Purdue University

Alignment plays a critical role in training high-quality Large Language Models (LLMs). It guides LLMs to generate content that matches human values and goals, and also serves as a safeguard mechanism to prevent LLMs from being exploited by malicious users to create inappropriate responses or spread misinformation. We present a theoretical framework showing that popular LLM alignment methods—including RLHF and

its variants—can be viewed as divergence estimators between aligned (safe or preferred) and unaligned (harmful or less-preferred) distributions. This perspective explains the separation phenomenon in the latent space of aligned LLMs between safe and harmful prompts. This general divergence framework inspires a novel KL divergence-based alignment method (KLDO), whose empirical effectiveness is validated via simulation. We further demonstrate that using compliance–refusal datasets, rather than standard preference-based datasets, yields stronger separation and improved safety alignment. Finally, to quantify the separation effect, we propose a distance-based metric in the prompt representation space, which also acts as a statistically significant indicator for model safety.

Transformers Are Born Biased: Structural Inductive Biases and Their Practical Consequences

Li Siquan, Tong Yao, Wang Haonan, ⬩ Tianyang Hu

The Chinese University of Hong Kong, Shenzhen, National University of Singapore, National University of Singapore, The Chinese University of Hong Kong, Shenzhen

While transformers are typically assumed to be "blank slates" at random initialization, we demonstrate that untrained models exhibit systematic structural biases, including extreme token preferences. We provide a mechanistic explanation by identifying a contraction of token representations along initialization-dependent directions, which is driven by the interaction of asymmetric MLP activations and self-attention value aggregation. We show these biases persist throughout training, forming a stable model identity that enables SeedPrint, a method to fingerprint LLMs by their random initialization. Finally, we establish that attention mechanism's intra-sequence contraction is causally linked to the attention-sink phenomenon, providing a principled explanation for sink emergence and a pathway for systematic control.

Theoretical Modeling of Large Language Model Self-Improvement Training Dynamics Through Solver-Verifier Gap

Yifan Sun, Yushan Liang, Zhen Zhang, Xin Liu, ⬩ Jiaye Teng

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

Self-improvement is a significant techniques within the realm of large language model (LLM), aiming to enhance the LLM performance without relying on external data. Despite its significance, generally how LLM performances evolve during the self-improvement process remains underexplored. In this paper, we theoretically model the training dynamics of self-improvement via the concept of solver-verifier gap. This is inspired by the conjecture that the performance enhancement of self-improvement stems from the gap between LLM's solver capability and verifier capability. Based on the theoretical framework, we further show how to model the entire training trajectory. This framework allows quantifying the capability limit of self-improvement by fitting the theoretical model to the experiment results. We validate the effectiveness of the theoretical framework on various LLMs and datasets. Beyond self-improvement, we extend our analysis to investigate how external data influences these dynamics within the framework. Notably, we find that under limited external data regimes, such external data can be utilized at any stage without significantly affecting final performances, which accords with the empirical observations.

Qualified LLM Judge Aggregation with Calibrated Uncertainty

Yunan Wu

Tsinghua University

As LLM-generated artifacts increasingly influence high-stakes decisions, trustworthy evaluation requires reliable estimates of how humans would judge those artifacts and when automatic judgments are uncertain. LLM-as-a-judge systems offer a scalable alternative to expensive human annotation, yet individual LLM judges may be weakly aligned with human scores, predictable from superficial features rather than content, or systematically biased. In this work, we introduce a framework that turns multiple LLM judges into a human-calibrated probabilistic scoring system. We first use double/debiased machine learning to test whether each judge contributes genuine human-relevant signal beyond observed covariates. Qualified judges are then calibrated to human-label probabilities with a PIE-Bridge model and combined using loss-based weights. This calibrated probability distribution provides uncertainty estimates that allow the system to automatically score confident examples while routing uncertain cases to human annotation. Experiments on benchmark LLM-judging tasks show that qualified calibrated aggregation improves probabilistic accuracy over the best individual judge and yields useful uncertainty quantification for human-in-the-loop evaluation.

S131: Statistical Frontiers in Spatial Omics

Intelligent integration methods for single-cell chromatin accessibility sequencing data

⬩*Heyang Hua, Shengquan Chen*

Nankai University, Nankai University

Single-cell chromatin accessibility sequencing (scATAC-seq) provides critical insights into gene regulatory mechanisms, but

integrating multi-source data faces significant challenges from batch effects. To address the common limitations of existing methods—namely, the loss of biological heterogeneity due to over-correction and compromised downstream reliability from neglecting original feature space correction—this presentation introduces Fountain, a deep learning model based on regularized barycentric mapping. By incorporating intra-batch geometric information as a regularized constraint for inter-batch mapping, Fountain achieves precise cross-batch alignment while preserving biological heterogeneity, and enables online integration of new data without retraining. Furthermore, Fountain provides batch-corrected features in the original dimensional space, significantly enhancing cellular heterogeneity characterization and demonstrating superior biological interpretability in downstream tasks such as cell-type-specific heritability analysis.

Benchmarking Cell-Type-Specific Spatially Variable Gene Detection Methods Using a Realistic and Decomposable Simulation Framework

Weiqi Li, ⬩Xinzhou Ge, Yuan Jiang

Oregon State University, Oregon State University, Oregon State University

Identifying spatially variable genes within individual cell types is essential for characterizing spatially organized cell states. Several computational methods have been developed for identifying cell-type-specific spatially variable genes (ctSVGs), but their relative performance and practical utility under realistic biological complexity remain largely unknown. To address this gap, we present the first systematic benchmark study of all five existing ctSVG detection methods---CELINA, STANCE, C-SIDE, CTSV and spVC---using an integrated evaluation framework that combines idealized simulations, Xenium-based realistic simulations, and a decomposition-based diagnostic analysis. Across realistic datasets generated on various tissue types, all methods experienced sharp declines in detection accuracy and substantial inflation of false discoveries compared to idealized simulations. To explain this failure, we developed a new simulation framework that decomposes the “realness” of the realistic simulation into interpretable biological and technical components, enabling us to attribute method-specific performance losses to specific components. Together, our results show that no single method dominates across detection accuracy, scalability and usability, and we further clarify why current ctSVG methods fall short in realistic settings. We summarize these tradeoffs into a practical user guide to support method selection and highlight key challenges in developing robust, scalable ctSVG detection tools for real spatial transcriptomics data.

Conditional Semi-Distance Correlation: Measuring Conditional Dependence in Spatial Transcriptomics

⬩*Ziqing Guo, Yang Feng, Jingyi Jessica Li, Lucy Xia*

Hong Kong University of Science and Technology, New York University, Fred Hutchinson Cancer Center, Hong Kong University of Science and Technology

Measuring conditional dependence is crucial in various fields, including genetic association studies and graphical models. In spatial transcriptomics, one of the focuses is on elucidating the relationships between gene expression levels (continuous variables) and spatially-related covariates, such as spatial locations, brain layers, and cell types (categorical variables), conditioning on other factors. We introduce a novel measure of conditional dependence that assesses the relationship between a categorical random variable and a potentially high-dimensional random vector, conditioned on another random vector. This measure is based on semi-distance correlation and extends the concept of conditional distance correlation to incorporate categorical variables. Importantly, it serves as a general conditional dependence metric, unrestricted by linear or monotonic relationships. We further develop a corresponding conditional independence test, which employs a local bootstrap procedure to efficiently compute pvalues. In addition, we propose a screening method based on this metric and establish its sure screening property. Simulation studies demonstrate the superiority of the proposed procedures. We apply the proposed test to spatial transcriptomics data and identify genes whose expression exhibits relatively strong conditional dependence on cell type given spatial location.

MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning

⬩*Yuying Xie, Jingru Guo, Yan Cui, Hongzhi Wen, Rob Yang*

Michigan State University, Michigan State University, CartaBio Inc, Cartabio Inc, Cartabio Inc

While spatial profiling technologies offer high-resolution data, extracting meaningful spatial structures remains labor-intensive and reliant on domain-heavy manual annotation. We propose MARVEL, a supervised graph contrastive learning framework that embeds local microenvironments into a continuous representation space. By leveraging partial annotations as strong positives, MARVEL mitigates false-positive issues inherent in unsupervised methods. We demonstrate that MARVEL outperforms existing tools in three key tasks: transductive annotation, inductive querying, and the identification of novel microenvironments across tissue slices.

S140: Statistical Methods and Applications

On Novel Applications of Classical Linear Regression in Bio-

medical Research

Zhezhen Jin

Columbia University

In this talk, I will present a working model based on classical linear regression model to study the effect of exposure mixture on the continuous health outcomes. The approach can estimate the parameter for overall effect and to test specific hypotheses, including two tests for detecting the overall effect and one test for detecting unequal weights when the overall effect is evident.

Meta- vs. Mega-Analysis: A Long-Standing Debate in Omics Data Integration

Jian Zou

Chongqing Medical University

Omics data integration is often framed as a choice between meta-analysis and mega-analysis, but the better strategy depends on the scientific question and data structure. In this talk, we examine this debate through two case studies. For meta-analysis, we present Mutual Information Concordance Analysis (MICA), a method for detecting multi-class concordant biomarkers across studies. MICA captures concordance shared by all or only some studies and supports post hoc identification of the genes and studies driving the signal. In a mouse metabolism application, it identified biologically relevant genes missed by conventional methods. For mega-analysis, we evaluate harmonization methods for cancer miRNA-seq data in clustering and classification. We show that harmonization strongly affects downstream performance and that no single method is optimal across settings. Rather than asking which paradigm is universally superior, we show that effective omics integration depends on selecting the strategy that best preserves biological signal, controls technical heterogeneity, and matches the downstream analytic goal.

Preserving Rare Features in Big Data Regression: Balanced Subsampling

Qingkai Dong, Jun Yan, Qiang Zhang, ⬩ HaiYing Wang

University of Connecticut, University of Connecticut, Wills Eye Hospital, University of Connecticut

Rare binary covariates, or rare features, pose significant challenges for statistical modeling in both classification and regression, including numerical instability and slow convergence. These challenges are often exacerbated in large-scale datasets, where subsampling is used for computational feasibility. Unfortunately, existing subsampling methods frequently fail to retain rare features, thereby worsening modeling difficulties. This paper develops solutions for these deficiencies in regression models. We first show that the estimation efficiency for rare feature coefficients is determined by the relatively small number of rare observations, rather than by the massive full-data sample size. Building on this insight, we propose a rareness-aware extension of the classical optimal subsampling method that adjusts sampling probabilities to preserve rare feature information. We further introduce a balanced subsampling method that quantifies rarity through a balance score, ensuring sufficient representation of rare features without relying on pilots. The balanced method can also replace traditional pilots in classical optimal subsampling, thereby enhancing robustness. Through theoretical analysis, simulation studies, and real-data applications, we demonstrate that the proposed methods dramatically improve both model-fitting robustness and estimation efficiency compared with existing subsampling approaches.

Simplifying Inputs of Machine Learning-Based Risk Calculators to Improve Clinical Usability

Zhe Lu, Kenneth Liao, Zhiyu Kang, Linwan Xu, Cindy Im, ⬩ Yan Yuan

University of Alberta, University of Minnesota, University of Minnesota, University of Alberta, University of Minnesota, University of Alberta

BACKGROUND: We published a machine learning algorithm Extreme Gradient Boosting (XGBoost) based algorithm for age-specific basal cell carcinoma (BCC) risk prediction in childhood cancer survivors. The algorithm currently available on the Childhood Cancer Survivor Study (CCSS) website needs 19 different clinical inputs. From the user perspective, automatic variable selection by XGBoost results in a model with high administrative burden because data for all considered inputs is needed. To improve the usability of these algorithms, we aim to leverage the performance benefits of machine learning algorithms and simplify the number of model inputs with little if any sacrifice in prediction accuracy.

METHODS: We investigated a number of input selection approaches for the XGBoost algorithm, the best performing algorithm in our BCC study. These approaches are L1-penalty, Elbow, SHAP-select, and Tree-structured parzen estimator (TPE). The manual Elbow method sequentially removes the input with smallest SHAP value (i.e. least important) until the reduced set shows worse performance than the previous larger set in all three evaluation metrics: discrimination (AUROC), precision (AUPRC), and overall performance (sBrS). The SHAP-select uses logistic regression of the outcome on the SHAP values of the model inputs, on the test data, and uses the regression coefficient signs and p-values to conduct inputs selection. TPE treats the inputs as a boolean list which is a part of the model hyperparameters to be optimized with respect to a user-specified objective function (i.e. sBrS). A simulation study was conducted where the simulated event-to-time outcome T depended on 5 key inputs, two of which were rare exposures (9% and 3%). Five noise inputs were added

to assess the ability of these four input selection approaches. We then applied the promising approaches to the BCC risk prediction algorithm using the CCSS data and evaluated the model performance on an external St. Jude Lifetime Cohort (SJLIFE) data, against the XGBoost algorithm using the full set of 19 inputs.

RESULTS: In 500 simulation datasets (n=5365, mean event rate=0.27), Elbow, SHAP-select and TPE kept on average 4.4, 5.0, and 6.35 inputs (of which 0.76, 0.22, and 1.50 were noise inputs), respectively. The L1-penalty performed poorly. In the BCC real-world example, Elbow, SHAP-select and TPE reduced the 19 inputs to 12, ##, and 14. The resulting reduced models had similar or better prediction performance than the full XGBoost model (AUROC 0.762 vs. 0.760; AUPRC 0.321 vs. 0.308; sBrS 0.079 vs. 0.074). Commonly dropped inputs included radiation therapy doses to the limbs, and specific chemotherapies (anthracyclines, platinum, corticosteroids).

CONCLUSION: Three input selection methods consistently drop noise inputs in simulation studies. In the use case, BCC risk prediction with reduced inputs by the three approaches outperformed the full-set XGBoost model in external validation data.

S148: Survival Analysis and Related Statistical Methods

Simultaneous Variable Selection and Estimation for a Partially Linear Cox Model

Tingting Cai, Mengqi Xie, Tao Hu, Jianguo Sun

Capital Normal University, Capital Normal University, Capital Normal University, University of Missouri

We consider simultaneous variable selection and estimation for a deep neural network-based partially linear Cox model and propose a novel penalized approach. In particular, a two-step iterative algorithm is developed with the use of the minimum information criterion to ensure sparse estimation. The proposed method circumvents the curse of dimensionality while facilitating the interpretability of linear covariate effects on survival, and the algorithm greatly reduces the computational burden by avoiding the need to select the optimal tuning parameters that is usually required by many other popular penalties. The convergence rate and asymptotic properties of the resulting estimator are established along with the consistency of variable selection. The performance of the procedure is demonstrated through extensive simulation studies and an application to a myeloma dataset.

Probit Time-to-Event Regression for Misclassified Group Testing Data

Lijun Fang, Tao Hu, Shuwei Li, Lianming Wang, Christopher S. McMahan, Joshua M. Tebbs

Capital Normal University, Capital Normal University, Guangzhou University, University of South Carolina, Clemson University, University of South Carolina

Group testing has been used extensively to reduce screening costs in epidemiological studies involving low-prevalence diseases. This testing strategy involves combining specimens (e.g., blood, urine, swabs, etc.) from several individuals to form a pool and then testing the pooled specimen for infection. When the endpoint of interest is a time-to-event outcome, for example, the time until infection or disease, and pools are measured only once, the resulting data are called group-tested current status data (petito2016misclassified). In this paper, we propose a new type of regression analysis for these data using a semiparametric probit model, an alternative to the proportional hazards model in survival analysis. A sieve maximum likelihood estimation approach is developed that approximates the model's nonparametric nuisance function by using logarithmic monotone splines, and an efficient expectation-maximization algorithm is proposed. Asymptotic properties of the resulting estimators are investigated by using empirical process techniques and sieve estimation theory. Numerical results from simulation studies suggest our estimation methods perform nominally, even when pools are possibly misclassified due to assay error, and can outperform individual testing when the number of assays (tests) is fixed. We illustrate our work by estimating a time-to-event regression model for chlamydial infection using group testing data from a large public health laboratory in Iowa.

Distance Weighted Learning for Individualized Treatment Rules

Wenying Wang, Jichang Yu, Yuanshan Wu

Zhongnan University of Economics and Law

One common approach to estimating Individualized Treatment Rules (ITRs) is the use of convex surrogate functions, whereby the original indicator function is replaced by a convex alternative, most frequently the hinge loss. This method, employed in approaches such as outcome weighted learning and augmented outcome-weighted learning, classifies observations based solely on a margin-based criterion. To enable more data points to exert a direct influence to on the estimation of the decision function, we propose adopting the distance weighted discrimination function as the convex surrogate in this framework, which also leads to improved computational efficiency. We establish the Fisher consistency, convergence properties, and risk bounds of the estimated ITR. The effectiveness of the proposed method is validated through extensive simulations and a real world case analysis, demonstrating improved performance over existing ITR

estimation approaches.

Fusion-Based Sparse Estimation for Current Status Data with Heterogeneous Auxiliary Information

◆ *Yijun Wang, Jie Li*

Zhejiang Gongshang University

Integrating auxiliary summary information from external studies into the analysis of internal individual-level data has become an established strategy for improving estimation efficiency. This paper proposes a doubly penalized Generalized Method of Moments (GMM) framework based on the Minimum Information Criterion (MIC) for the additive hazards model with current status data. The proposed framework simultaneously achieves sparse estimation of high-dimensional regression parameters and identifies potential incompatibility (heterogeneity) between internal and external data by introducing a heterogeneity parameter. Theoretically, the consistency and asymptotic normality of the estimators are established, with explicit formulas derived for the asymptotic variance-covariance matrix. Extensive simulation studies are conducted and demonstrate that the proposed method outperforms internal-data-only methods in statistical efficiency and, in the presence of heterogeneity, significantly reduces estimation bias and standard errors. Finally, the practical value of the method is illustrated through an analysis of hypertension risk factors using data from the National Health and Nutrition Examination Survey (NHANES). The proposed method successfully identifies key risk factors—including waist circumference, glycohemoglobin, uric acid, and creatinine—and detects heterogeneity within the Black, high school or less, and college graduate subgroups. The results demonstrate the accuracy of the proposed method, while sensitivity analysis further confirms its stability and superiority.

S106: Recent Advances in Sequential Methodologies

Intelligence-Driven Quality Analytics: Integrating Generative AI with Statistical Process Control

Fugee Tsung

Hong Kong University of Science and Technology

Despite the rapid rise of Generative AI, 95% of industrial pilots fail. This keynote addresses this challenge by exploring the critical transition from quality engineering to industrial intelligence, from traditional industrial informatics to intelligence-driven decision-making. Drawing on cutting-edge research from HKUST's Quality Lab and Triple-I Institute, we will examine the practical applications and engineering limitations of Industrial AI, highlighting advancements in anomaly detection, multi-modal analytics, and physics-informed models. Attendees will gain actionable strategies to ground AI in industrial reality, build essential safety guardrails, and drive high-value transformations, ultimately challenging them to define their role in the future of intelligent industrial ecosystems.

Nearest-Neighbor Methods for Sequential Change-Point Detection in High-Dimensional and Non-Euclidean Data

Hao Chen

University of California, Davis

A central problem in sequential analysis is to detect distributional departures in a data stream with small detection delay, subject to a prescribed false-alarm constraint. This problem is increasingly challenging in modern applications where the observations may be high-dimensional vectors or non-Euclidean objects, such as networks or images, and where fully specified likelihood models are often unavailable. This talk presents a nonparametric framework for sequential change-point detection based on nearest-neighbor information. By exploiting local similarity relationships among recent observations, the proposed methods provide a flexible approach to online monitoring for complex data streams. I will discuss key methodological ideas, theoretical calibration for false-alarm control, and extensions that improve detection power for a broad class of alternatives. Numerical studies and real-data examples demonstrate the effectiveness of the approach.

Sequential Multistream and Multihypothesis Testing With Asynchronous Decisions and Generalized Error Metrics

Yiming Xing

Tongji University

The problem of simultaneously testing multiple hypotheses about the distribution of each of multiple independent, sequentially observed data streams is considered. The decisions for the streams can be made at different times, using available data from all streams, and a stream is no longer observed once a decision about it has been made. Besides, it is required to control flexible error metrics that do not penalize the incorrect selection of a hypothesis if the number of streams where this occurs does not exceed a certain threshold, which include the generalized familywise error rates as a typical example. In this general framework, unlike previous works it is impossible to minimize, even asymptotically, the expected sample size simultaneously in all streams. Thus, we aim to minimize the expected total sampling cost, where the cost per observation may vary across streams. A universal asymptotic lower bound is established for this quantity as the target levels of the error metric go to zero. Moreover, a novel testing procedure is proposed, analyzed, and shown to achieve this lower bound under certain conditions. This asymptotic optimality theory is established for general parametric composite hypotheses and

allows for temporal dependence. In addition, a novel synchronous procedure and a novel family of decentralized procedures are proposed and compared with the optimal one, which quantify the benefit of allowing for asynchronous decisions and the existence of a global decision maker.

Optimal Bounded-Influence Procedure for Robust Sequential Change-Point Detection

Ruizhi Zhang

University of Georgia

We study the problem of robust sequential change-point detection under Huber's gross error model. We incorporate ideas of the influence function from the classical offline robust statistics literature and propose a new definition called the false alarm influence function to quantify the robustness of general sequential change detection procedures. Then, we propose a general family of robust detection procedures and derive their false alarm influence function. The result shows that they all have bounded influence functions, which implies that any arbitrary outlier will have only a limited impact on their detection performance. Furthermore, we construct the optimal robust bounded-influence procedure in that general family that has the smallest detection delay subject to the false alarm rate constraint and the false alarm influence function constraint. It turns out the optimal procedure is based on the truncated likelihood ratio statistics.

S012: Advances in Machine Learning Methods for Survival and Event-Time Data

DH-AFT: Distributed, Heterogeneity-Aware AFT Models for High-Dimensional Multi-Center Survival Data

Hong Wang

Central South University

Accelerated failure time (AFT) models provide a flexible alternative to proportional hazards models for analysing time-to-event data, particularly in biomedical and genomic applications with complex survival mechanisms. In modern multi-center studies, however, centralised analysis of high-dimensional survival data is often hindered by privacy regulations, communication constraints, and computational burdens. To address these challenges, we propose DH-AFT, a distributed and heterogeneity-aware framework for high-dimensional AFT models in multi-center survival studies. The method constructs a global quadratic surrogate loss using only center-level summary statistics from local Lasso estimators, and combines a group sparsity penalty with a fusion penalty on site-specific regression coefficients to enable simultaneous variable selection and adaptive clustering of centers. Optimization is carried out via a proximal ADMM algorithm coupled with FISTA, and we establish oracle-type theoretical guarantees showing that, under mild conditions, DH-AFT asymptotically attains the performance of the oracle estimator. Extensive simulations and a real-data application to multi-center ovarian cancer gene expression cohorts indicate that DH-AFT achieves accurate and stable estimation and prediction while relying only on summary-level information.

Improving Prediction for Interval-Censored Data Using Auxiliary Predictive Information

Jingyi Hu, ⋆Yu Gu

University of Hong Kong, University of Hong Kong

Accurate prediction for interval-censored survival outcomes is particularly challenging when censoring intervals are wide and follow-up is limited, as is often the case in studies of chronic diseases. Although auxiliary information from source studies may improve prediction in a target study, existing transfer learning methods typically impose restrictive assumptions on model or parameter similarity, or require access to individual-level source data. We propose a novel transfer learning method for interval-censored data that allows arbitrary source model classes and avoids sharing source data. Our approach transfers predictive information from source studies through a carefully designed penalty, which circumvents computational challenges via a simple EM algorithm. When multiple source studies are available and source informativeness is unknown, we further develop a data-driven procedure that is robust to negative transfer. Theoretical analysis shows that the proposed predictor attains a faster convergence rate than the target-only predictor whenever at least one source study is sufficiently informative. Extensive simulation studies and an application to data from the Alzheimer's Disease Neuroimaging Initiative demonstrate the effectiveness of our approach.

Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

⋆*Xiangbin Hu, Wen Su, Zhisheng Ye, Xingqiu Zhao*

Beijing Institute of Technology, City University of Hong Kong, National University of Singapore, The Hong Kong Polytechnic University

Personalized treatment regimes that account for individual characteristics offer substantial potential to minimize treatment risks and enhance patient outcomes. Existing methods for estimating individualized treatment rules in multi-arm settings often face limitations due to model misspecification. To overcome the challenges, we propose a novel conditional generative learning framework, CG-Learning, which employs a Wasserstein generative adversarial network to estimate optimal decision

rules that minimize a specified risk measure in multi-armed treatment scenarios. By modeling the conditional distribution of rewards through conditional generative adversarial networks, our approach avoids restrictive structural assumptions. We establish key theoretical properties of the resulting decision rules, including nonasymptotic bounds on regret and mis-assignment probabilities. We show that these bounds can be tightened when the covariate space has a low Minkowski dimension. Through comprehensive simulations and an application to data from the AIDS Clinical Trials Group, we demonstrate that CG-Learning outperforms existing benchmarks.

S002: Advanced Methods for Complex Data Analysis

Revisiting the Alpha-Folding Multivariate Normal Distribution

Xiaoke Qin, Saman Muthukumarana, ◆ Kevin McGregor

University of Manitoba, University of Manitoba, University of Manitoba

We revisit the alpha-folding multivariate normal distribution for compositional data analysis under two inferential frameworks. We express the model as a log-sum-exp representation, which highlights how the two branches arise from a many-to-one folding mechanism and their Jacobian terms, without a finite-mixture interpretation. For frequentist inference, we propose a practical maximum-likelihood method based on fixed-point estimation and profile likelihood, and construct likelihood-ratio confidence intervals for alpha. We additionally propose a Bayesian framework that targets the marginal posterior of alpha using a Metropolis-within-Gibbs sampler that alternates between updates of alpha and the Gaussian parameters, motivated by the strong coupling between them and the potential sensitivity of profile optimization to initialization. Through simulation studies and real-data analysis, we compare the two approaches in terms of estimation accuracy, numerical stability, and uncertainty quantification.

Bayesian Approaches for Multi-Class Fault Detection Classification

Saman Muthukumarana

University of Manitoba

We present a Bayesian and nonparametric Bayesian framework for multi-class classification of leakage levels in hydraulic actuators using pressure signal data. A multinomial logistic regression model is constructed using feature extraction, with Dirichlet Process and Gaussian Process priors to enable flexible, data-adaptive structure. An Infinite Hidden Markov Model (IHMM) is also employed to directly model temporal dependencies in raw signal sequences without feature extraction. Bayesian inference is performed using Markov Chain Monte Carlo for the regression-based models and collapsed Gibbs sampling for the IHMM. Leakage-level predictions are obtained from posterior predictive distributions, providing full uncertainty quantification. Model performance and hyperparameters are assessed through k-fold cross-validation using multiple classification metrics.

Thompson Sampling with Bayesian Additive Regression Trees

Ruizhe Deng, Bibhas Chakraborty, Ran Chen, ◆ Yan Shuo Tan

National University of Singapore, Duke-NUS Medical School, Washington University in St. Louis, National University of Singapore

Contextual bandits are a core technology for personalized mobile health interventions, where decision-making requires adapting to complex, non-linear user behaviors. While Thompson Sampling (TS) is a preferred strategy for these problems, its performance hinges on the quality of the underlying reward model. Standard linear models suffer from high bias, while neural network approaches are often brittle and difficult to tune in online settings. Conversely, tree ensembles dominate tabular data prediction but typically rely on heuristic uncertainty quantification, lacking a principled probabilistic basis for TS. We propose Bayesian Forest Thompson Sampling (BFTS), the first contextual bandit algorithm to integrate Bayesian Additive Regression Trees (BART), a fully probabilistic sum-of-trees model, directly into the exploration loop. We prove that BFTS is theoretically sound, deriving an information-theoretic Bayesian regret bound. As a complementary result, we establish frequentist minimax optimality for a "feel-good" variant, confirming the structural suitability of BART priors for non-parametric bandits. Empirically, BFTS achieves state-of-the-art regret on tabular benchmarks with near-nominal uncertainty calibration. Furthermore, in an offline policy evaluation on the Drink Less micro-randomized trial, BFTS improves engagement rates by over 30% compared to the deployed policy, demonstrating its practical effectiveness for behavioral interventions.

Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach

Lei Ding

University of Manitoba

As generative large language models (LLMs) such as ChatGPT gain widespread adoption in various domains, their potential to propagate and amplify social biases, particularly in high-stakes areas such as the labor market, has become a pressing concern. AI algorithms are not only widely used in the selection of job applicants, individual job seekers may also make use of generative LLMs to help develop their job application materials. Against this

backdrop, this research builds on a novel experimental design to examine social biases within ChatGPT-generated job applications in response to real job advertisements. By simulating the process of job application creation, we examine the language patterns and biases that emerge when the model is prompted with diverse job postings. Notably, we present a novel bias evaluation framework based on Masked Language Models to quantitatively assess social bias based on validated inventories of social cues/words, enabling a systematic analysis of the language used. Our findings show that the increasing adoption of generative AI, not only by employers but also increasingly by individual job seekers, can reinforce and exacerbate gender and social inequalities in the labor market through the use of biased and gendered language.

S092: Recent Advances for the Analysis of Complex Data

Functional feature screening with false discovery rate control for ultrahigh-dimensional longitudinal data

Yexin Zhang

Wuhan University

In many modern biomedical and public health studies, ultrahigh-dimensional longitudinal data are often collected through repeated measurements on subjects. Developing methods that can efficiently screen for relevant covariates while rigorously controlling false discovery rate (FDR) is essential, yet this remains a challenging problem for such data. In this paper, we propose a novel method that extends Pearson's correlation to a functional correlation between time-dependent covariates and the response process to perform feature screening. By adopting the data splitting strategy, we construct a series of symmetric statistics and then utilize the symmetry property to determine a data-driven threshold to screen out irrelevant covariates, so as to achieve the FDR control of the screened model. The proposed method is computationally efficient and, under some mild conditions, possesses both the sure screening property and FDR control simultaneously. Simulation studies under various settings demonstrate its promising screening and FDR control performance. Two real-data applications further validate its practical effectiveness.

Data-driven Newsvendor Decision-making: Newsvendor Representation Transfer Learning

Baihua He

University of Science and Technology of China

We study the data-driven newsvendor problem in situations where historical demand data and associated feature information are available. To address this problem, we leverage information from similar data sources and employ transfer learning to improve decision-making and facilitate statistical inference on the parameter of primary interest in the target domain. Within the framework of a semiparametric regression model, we assume the existence of a shared feature representation for confounding effects across different tasks. By learning this shared representation from various source domains, we can effectively transfer it to the target domain. This approach enables both accurate decision-making and interpretability in the target domain. We establish sufficient conditions for model identifiability, derive a finite-sample performance bound for the cost function, and prove that the estimator of the parameter of primary interest in the target model is consistent and asymptotically normal. Through simulation studies, we demonstrate the superiority of our method, and we further illustrate its practical applications in inventory decision-making for bike-sharing systems.

Regression analysis of recurrent gap times with informative observation and censoring times

⋆*Shu-Hui Chang, Hsin-Yang Lu*

Taiwan University, Taiwan University

In longitudinal studies of chronic diseases, the gap/remission time from a clinical intervention of one recurrent event to the next event is a natural outcome of interest for clinical evaluation. However, the exact time of recurrent event is not effortless to be determined, and the status of recurrent event is observed only at clinical visit or checkup. The arrangement and numbers of clinical visits may depend on the status of recurrent disease and relevant clinical features. Such outcome-dependent visit process is subject to censoring such as loss to follow-up. Consider semiparametric regression models for recurrent gap times without specifying the within-subject correlation between recurrent gap times. A weighted estimation method for model parameters is developed using the inverse intensity of visit and the inverse probability of censoring to handle the dependent visit and censoring processes, respectively. The asymptotic properties of the proposed method are studied. Simulation studies and a data example are used to demonstrate the proposed method.

S014: Advances in Modern High-dimensional Statistics

A Frequency-Domain Approach for Integrating Multiple Functional Time Series

Zerui Guo, Jianbin Tan, ⋆*Hui Huang*

Sun Yat-sen University, Duke University, Renmin University of China

Integrative analysis of multivariate functional time series (MFTS) is both critical and challenging across many scientific domains. Such data often exhibit complex multi-way dependencies arising from within-curve structures, temporal correlations across curves and cross-subject interactions, underscoring the need for efficient methods that can jointly capture these dependencies and support accurate downstream analyses. In this work, we propose a novel frequency-domain framework based on a marginal dynamic Karhunen–Loève expansion. The key idea is to integrate individual spectral densities of the MFTS to construct a marginal spectral operator, whose eigenfunctions yield optimal functional filters. These filters transform complex functional observations into a structured multivariate time series representation, providing a powerful foundation for joint modelling and estimation. Through extensive simulation studies, we demonstrate the superior performance of the proposed approach. We further validate its practical utility through an application to the imputation and forecasting of air pollutant concentration trajectories in China.

Triple-Dyad Ratio Estimation for the p1 Model

Qunqiang Feng, Yaru Tian, ⋆ Ting Yan

University of Science and Technology of China, University of Science and Technology of China, Central China Normal University

Although the p1 model was proposed 40 years ago, little progress has been made to address asymptotic theories in this model, that is, neither consistency of the maximum likelihood estimator (MLE) nor other parameter estimation with statistical guarantees is understood. This problem has been acknowledged as a long-standing open problem. To address it, we propose a novel parametric estimation method based on the ratios of the sum of a sequence of triple-dyad indicators to another one, where a triple-dyad indicator means the product of three dyad indicators. Our proposed estimators, called triple-dyad ratio estimator, have explicit expressions and can be scaled to very large networks with millions of nodes. We establish the consistency and asymptotic normality of the triple-dyad ratio estimator when the number of nodes reaches infinity. Based on the asymptotic results, we develop a test statistic for testing whether is a reciprocity effect in directed networks. The estimators for the density and reciprocity parameters contain bias terms, where analytical bias correction formulas are proposed to make valid inference. Numerical studies demonstrate the findings of our theories and show that the estimator is comparable to the MLE in large networks.

Covariate-Dependent Latent Gaussian Graphical Models for Mixed-Type Data

Yuyang Liu, Liuyi Lan, ⋆ Shan Luo, Yiwang Zhou, Chenlei Leng

Shanghai University of International Business and Economics, Shanghai Jiao Tong University, Shanghai Jiao Tong University, St. Jude Children's Research Hospital, Hong Kong Polytechnic University

We introduce a general framework for estimating covariate-dependent graphical models for multivariate data with mixed response types, including categorical, ordinal, and continuous variables. Our method, based on a covariate-dependent latent Gaussian graphical model (CD-LGGM), enables individualized network estimation that adapts to subject-specific covariates. We develop a penalized expectation-maximization algorithm for high-dimensional settings, establish theoretical guarantees for graph recovery and parameter estimation, and demonstrate superior performance over existing methods in simulations. Application to the Cleveland Heart Disease dataset illustrates the utility of CD-LGGM for constructing personalized clinical phenotype networks. The framework is broadly applicable to network estimation with heterogeneous and covariate-dependent data.

Decoupling Regularization and Privacy in Differentially Private Ridge Regression and ERM

⋆ *Wanjie Wang, Tathagata Banjernee*

National University of Singapore, National University of Singapore

We study ridge regression and ridge-regularized empirical risk minimization (ERM) under (\varepsilon, \delta)-differential privacy via output perturbation. In classical private ERM, the ridge parameter simultaneously controls statistical regularization and the estimator's global sensitivity. Larger regularization reduces the DP noise scale but increases bias. So choosing the tuning parameter becomes a privacy--accuracy bottleneck. We propose a framework that makes these two roles explicit by decoupling regularization into (i) a statistical penalty α, defining the target ridge/ERM solution, and (ii) a privacy-stabilization parameter c, used only to enforce a curvature floor and hence a tight sensitivity bound. We apply this framework to ridge regression, where is used to boost the minimum eigenvalue of the empirical Gram matrix. We derive an explicit bias--variance--DP-variance risk decomposition and characterize optimal in several regimes, yielding sharp tuning guidance and improved accuracy relative to single-parameter regularization. Finally, we extend the same decoupling principle to general ridge-regularized ERM. We support the theory with simulations.

S022: Advances in Survival Analysis and Causal Inference

Inference and Variable Selection for Two-Phase Studies with High-Dimensional Covariates

Haoyang Wang, Qingning Zhou, ⋆ *Kin Yau Wong*

The Hong Kong Polytechnic University, The University of North Carolina at Charlotte, The Hong Kong Polytechnic University

The two-phase study design is widely used to improve estimation efficiency and reduce cost. In many two-phase studies, the outcome and inexpensive covariates are obtained on all subjects in Phase I, while expensive covariates are measured only on a subset of subjects in Phase II. As a result, regression analysis of two-phase studies faces a missing data problem. In this project, we consider two-phase studies with high-dimensional covariates, where one faces a more challenging high-dimensional missing data problem. For this problem, complete-case analyses are generally inefficient, while imputation or likelihood-based methods require a model for the missing covariates, which is almost impossible to correctly specify. To overcome these limitations, we propose a two-step estimation method that refines a complete-data estimator by incorporating the incomplete data, and auxiliary information if available, to improve efficiency. This method does not require modeling the distribution of the missing covariates, and the resulting estimator is guaranteed to be at least as efficient as the complete-data estimator. We establish theoretical properties of the proposed method, including estimation consistency, inference validity, and variable selection consistency. We evaluate the performance of the proposed method via simulation studies and provide an application to a major cancer study.

Semiparametric Causal Inference for Right-Censored Outcomes with Many Weak Invalid Instruments

Qiushi Bu, ⋆ *Wen Su, Zhonghua Liu, Xingqiu Zhao*

City University of Hong Kong, City University of Hong Kong, Columbia University, The Hong Kong Polytechnic University

We propose a semiparametric framework for causal inference with right-censored survival outcomes and many weak invalid instruments, motivated by Mendelian randomization in biobank studies where classical methods may fail. We adopt an accelerated failure time model and construct a moment condition based on augmented inverse probability of censoring weighting, incorporating both uncensored and censored observations. Under a heteroscedasticity-based condition on the treatment model, we establish point identification of the causal effect despite censoring and invalid instruments. We propose GEL-NOW (Generalized Empirical Likelihood with Non-Orthogonal and Weak moments) for valid inference under these conditions. A divergent number of Neyman orthogonal nuisance functions is estimated using deep neural networks. A key challenge is that the conditional censoring distribution is a non-Neyman orthogonal nuisance, contributing to the first-order asymptotics of the estimator for the target causal effect parameter. We derive the asymptotic distribution and explicitly incorporate this additional uncertainty into the asymptotic variance formula. We also introduce a censoring-adjusted over-identification test that accounts for this variance component. Simulation studies and UK Biobank applications demonstrate the method's robustness and practical utility.

Deep Nonparametric Inference for Interval-Censored Data

⋆ *Qiang Wu, Wen Su, Shuangge Ma and Xingqiu Zhao*

The Hong Kong Polytechnic University

Interval-censored failure time data pose significant challenges in survival analysis, particularly when capturing complex covariate interactions. Traditional statistical approaches, such as the Cox proportional hazards model, are often limited by strict structural assumptions. Conversely, fully nonparametric methods, like interval-censored recursive forests, relax these assumptions but lack the capacity for rigorous statistical inference. To overcome these limitations, we propose a novel deep neural network-based approach to directly estimate the conditional cumulative hazard function. A key contribution of this work is establishing a rigorous theoretical foundation for deep survival models. Specifically, we derive nonasymptotic error bounds and establish nonparametric asymptotic normality, thereby facilitating valid statistical inference, a critical component often absent in deep learning applications. Furthermore, we develop novel procedures for two-sample testing and goodness-of-fit evaluation. Extensive simulation studies demonstrate that our proposed estimator achieves superior predictive accuracy and maintains high performance across diverse data-generating mechanisms compared to state-of-the-art methods. Finally, an analysis of the Atherosclerosis Risk in Communities study confirms its strong empirical performance.

S019: Advances in Statistical Methods for Clinical Trials: Longitudinal Data and Complex Data Mechanisms

Uncovering Hormone Trajectory Subtypes in the Menopausal Transition via Envelope–Regularized Multivariate Longitudinal Clustering

⋆ *Zihang Lu, Bei Jiang*

University of Toronto

Clustering multivariate longitudinal biomarker trajectories is often challenging due to nonlinear temporal dynamics, strong cross-biomarker dependence, and the need to disentangle shared system-level signals from biomarker-specific variation. Motivated by the Penn Ovarian Aging Study, we propose a Bayesian framework for clustering nonlinear multivariate longitudinal

hormone trajectories via matrix-variate envelope regularization. Our approach leverages envelope structure as a regularization mechanism to decompose matrix-variate subject-specific random effects into material and immaterial components, effectively removing variation irrelevant to clustering. This leads to improved statistical efficiency and interpretability. Applying the proposed method to longitudinal hormone data, we identify three clinically interpretable hormone trajectory subtypes during the menopausal transition, uncovering the heterogeneity in endocrine dynamics that are not captured by conventional methods.

Inverse Probability of Treatment Weighting for Estimation of Time Dependent Treatment Effects Under the Test-Negative Design

⬩ *Helen Bian, Cong Jiang, Denis DenisTalbot, Robert Platt, Mireille Schnitzer*

McGill University, Harvard University, Laval University, McGill University, University of Montreal

The test-negative design (TND) has been widely used for the rapid estimation of vaccine effectiveness against infectious diseases. The TND typically includes individuals with a common symptom profile who are receiving a laboratory test for an infection of interest. Among them, participants who test positive for the target infection are “cases” and those who test negative are “controls”. Existing statistical approaches for the TND have certain limitations in a dynamic longitudinal setting, where data can be periodically collected from different individuals over the study period.

First, time-dependent confounders may be influenced by previous vaccination and health status, such as previous infection, while also affecting the subsequent vaccination decisions. Thus, the causal relation of interest cannot be properly estimated using traditional covariate-adjusted models. Secondly, since individuals can test positive multiple times over the study period, past infections may alter immunity and create non-positivity for vaccination for the following time-period. Therefore, we propose a causal framework that accounts for the time-varying effects as well as changing risk sets, particularly in the context of the dynamic nature of infectious diseases. We propose the inverse probability of treatment weighting estimators of discrete-time hazard and hazard ratio, which can be identified using the TND samples. Sandwich variance estimators are derived for the proposed estimators. Simulation studies are used to show the performance of these estimators.

Disease Progression as a Common Driver of HRQoL Trajectories and Observation Schedules in Advanced Cancer Trials: A Progression-Stratified Pattern-Mixture Analysis

⬩ *Wensha Zhang, Wei Tu*

Queen’s Univerisity, Queen’s University

In advanced cancer trials, disease progression simultaneously shapes patients’ HRQoL experience, on-study assessment frequency, and post-progression follow-up density. When treatment duration differs between arms by design, as is especially pronounced in immunotherapy trials, patients in the longer-treatment arm naturally contribute denser HRQoL observations before progression. Standard mixed-effects models risk producing comparisons that reflect differences in observation schedules as much as true differences in patient experience. We propose a progression-stratified pattern-mixture framework that addresses this by defining analytic patterns according to progression-free survival, the common upstream clinical driver of both HRQoL trajectories and assessment processes. We apply this approach to HRQoL data from the CCTG IND227 trial in advanced mesothelioma. Patients are stratified into progression-defined groups using PFS quartiles, and spline-based mixed-effects models are fitted within each group and combined using arm-specific weights. As a complementary perspective, we also conduct terminal decline analyses indexed by time to death. The proposed framework reveals substantial heterogeneity in HRQoL trajectories that is not captured by standard models. Compared with conventional analyses, the pattern-mixture approach shows steeper late-phase declines and larger between-arm differences. The chemo-immunotherapy arm demonstrates consistently better HRQoL, with stronger effects under the proposed model. Overall, progression-stratified pattern-mixture models provide a clinically interpretable and practical approach for HRQoL analysis in settings with differential follow-up, particularly in immunotherapy trials.

Loss to Follow-Up in Phase III Cancer Trials: Patient-Level Predictors and Impact on Treatment Effect Estimates

⬩*Wei Tu, Nick Cruickshanks, Harriet Richardson*

Queen’s University and Canadian Cancer Trials Group, Queen’s University, Queen’s University and Canadian Cancer Trials Group

Loss to follow-up (LTFU) and withdrawal of consent (WOC) represent a source of uncertainty in randomized controlled trials with time-to-event endpoints, yet their patient-level determinants and consequences for treatment effect estimates remain incompletely understood in oncology. Using individual patient-level data from a large Phase III cancer trial, we examine characteristics associated with LTFU and WOC status, identifying demographic and clinical factors that differ between those lost or withdrawn and those completing follow-up, raising concern for potentially informative censoring. We then evaluate the sensitivity of the primary time-to-event treatment effect to a range of assumptions about the unknown outcomes of these patients, applying tipping point methodology to trace how the hazard

ratio and statistical conclusion evolve across scenarios from worst-case to best-case imputation. A key observation emerging from this analysis is that existing sensitivity frameworks treat all informatively censored patients as exchangeable, assigning equal weight to a patient lost at week four and one who withdrew after five years of follow-up — despite the fundamentally different amounts of information each has already contributed. We propose that a more principled approach should condition on both the reason and timing of censoring, and outline a sensitivity framework that does so, with the aim of providing a more transparent and clinically meaningful assessment of how dropout affects confidence in trial conclusions.

S029: Big Data and AI

Flexible Semiparametric Modeling With Application to Causal Inference

Xingqiu Zhao

The Hong Kong Polytechnic University

This paper proposes a flexible new framework for constructing Neyman-orthogonal scores in semiparametric models involving infinite-dimensional nuisance parameters. While locally robust estimation is vital for integrating machine learning into econometrics, deriving orthogonal scores for complex models remains a major challenge. We provide explicit construction strategies for broad classes of settings. The proposed framework ensures asymptotic normality of target parameter estimators in a way that does not depend on the method used to construct the nuisance parameter estimators, provided they are $o_p(n^{-1/4})$-consistent. We apply the proposed methodology to causal inference with a binary instrumental variable, developing a novel, robust estimator for treatment effects. Numerical studies demonstrate that our approach significantly outperforms naive alternatives in finite samples. An empirical application to the Oregon Health Insurance Experiment illustrates the framework's utility in providing robust causal evidence.

A Deep Learning Prognostic Model Derived from Whole-Transcriptome Data in Lung Adenocarcinoma

Hsuan Yu Chen

Institute of Statistical Sciences, Academia Sinica, Taiwan

Targeted therapies and immunotherapies have improved outcomes in lung adenocarcinoma. However, recurrence and limited survival remain major clinical challenges. We developed an image-based deep learning framework for prognosis prediction using whole-transcriptome data. Four cohorts were analyzed, including one RNA-seq training cohort (n=391) and three independent validation cohorts: TCGA (n=394), GSE68465 (n=443), and GSE13213 (n=117). Gene expression profiles were transformed into image representations through pixel encoding, and a convolutional neural network was trained using all genes without feature selection. The training cohort was divided into training (n=259), testing (n=62), and internal validation (n=70) subsets, followed by external validation across independent cohorts. The model consistently stratified patients into high- and low-risk groups. High-risk patients had significantly worse overall survival in the training cohort ($P<0.01$) and in TCGA ($P=0.001$), GSE68465 ($P<0.0001$), and GSE13213 ($P=0.003$). Prediction accuracies were 0.71, 0.81, and 0.92 in the training, testing, and internal validation subsets, respectively, and 0.70 across external cohorts. These findings demonstrate that image-based deep learning of whole-transcriptome data enables robust and platform-independent prognostic prediction without gene selection, supporting scalable applications in precision oncology.

Domain-Specific Nonparametric Regression for Domain Generalization

◆ *Siming Zheng, Yuanyuan Lin, Yong Zhou, Jian Huang*

Southeast University, The Chinese University of Hong Kong, East China Normal University, The Hong Kong Polytechnic University

We propose a domain-specific regression approach for domain generalization, taking into account possible heterogeneity among the datasets from different sources. In the proposed model, the domain-specific features are characterized through linear functionals of the marginal source distributions. The predictors are combined with the domain-specific linear functionals as inputs in the model. Using the source data sets, we estimate the domain-index function and the regression function nonparametrically based on neural network approximation. The resulting estimated domain-specific regression function can be used for prediction when future unlabelled data from a target domain arrives. The proposed method is shown to be consistent in the sense the cross-domain prediction error of the estimated domain-specific predictive function converges to that of the underlying true domain-specific predictive function. We also establish the convergence rates under suitable conditions. We demonstrate the performance of the proposed method through numerical experiments with simulated and real data. The numerical results show that, our method outperforms the naive pooling method and the nearest-domain-prediction method and our method is robust to model assumptions.

S040: Experimental Design and Optimization

A General (Non-Markovian) Framework for Covariate Adaptive Randomization: Achieving Balance While Eliminating the Shift

Hengjia Fang, ⋆ *Wei Ma*

Institute of Statistics and Big Data, Renmin University of China

Covariate adaptive randomization (CAR) procedures employ a sequential allocation mechanism. In most existing procedures, these mechanisms induce a Markovian structure on the imbalance vector process for covariates. Because verifying the petite set condition in the Markov chain theory depends on the structure of the state space, establishing the boundedness of this process in probability often requires arguments that differ across covariate types. Moreover, for CAR procedures under equal allocation, the imbalance in additional covariates not targeted by the CAR procedure is known to be asymptotically normal and centered at zero. However, under unequal allocation, the center of the asymptotic distribution shifts away from zero. This article addresses these issues in a general framework, where the allocation mechanism is allowed to depend on the entire history, violating the Markov property. Within this framework, we first establish a unified condition for the boundedness applicable to all covariate types, via a novel Lyapunov function argument. We then provide a rigorous probabilistic explanation for the existence of the shift. Moreover, within this framework, we establish the asymptotic distribution of the imbalance in additional covariates. Motivated by these results, we construct a novel CAR procedure with an adaptively updated allocation mechanism that balances covariates and eliminates the shift.

A New Discrepancy-Driven Framework for Uniform Factorial Designs with Application to Bayesian Structure Learning

Sumin Wang

School of Sciences, Hebei University of Technology

In modern scientific research, many experiments require consideration of both the doses and the order of addition of components. The uniformity of a design is crucial for ensuring space-filling properties, which enhance the accuracy and reliability of results. This paper introduces a new uniformity criterion, the order–quantitative discrepancy, for evaluating designs with order-of-addition and dosage factors. Based on discrete discrepancy and wrap-around L_2-discrepancy, OQD provides a comprehensive measure of design uniformity. We derive its lower bound to benchmark design optimality and propose Uniform Ordering Factorial Designs (UOFDs) that achieve this bound. Several construction methods for UOFDs are provided, offering practical guidelines to improve experimental efficiency and effectiveness. Furthermore, we apply the proposed design to the Bayesian structure learning and introduce an order-based Markov chain Monte Carlo (MCMC) algorithm, which can rapidly yield an approximate solution to the structure learning challenge.

Uniform Projection Designs Under the Stratified L2-Discrepancy

Sixu Liu, ⋆ *Yaping Wang*

Beijing Institute of Mathematical Sciences and Applications, East China Normal University

We study a uniform projection criterion for space-filling designs under the stratified L2-discrepancy. The criterion, denoted by ΦSD, is the average squared stratified L2-discrepancy over all two-dimensional projections. For Utype (n,m,s^p) designs, we derive an explicit formula for ΦSD in terms of row-pairwise weighted hierarchical distances, and we establish sharp lower and upper bounds with equality conditions. We further show that many known optimal constructions attain the lower bound of ΦSD, and that designs attaining the lower bound of the full stratified L2-discrepancy also attain the lower bound of ΦSD. The criterion can be evaluated in O(n2m) time, with a modest reduction in arithmetic operations compared with direct projection-wise evaluation. Numerical studies illustrate the theoretical results and show that ΦSD is effective for assessing low-dimensional projection uniformity.

Optimal Design for Online Sequential Multi-Factorial Experiments with Optional Stopping

Xin Liu, ⋆ *Jun Yu, Wenbiao Zhao, Rui Zeng*

Donghua University, Beijing Institute of Technology, Beijing Institute of Technology, Beijing Institute of Technology

Online sequential experimental designs are critical in product development and efficacy validation. Effective sequential designs should immediately suspend trials if it appears that no effect will be detected.

Otherwise, it should provide a precise estimation of treatment effects or the response surfaces.

This work introduces an adaptive framework for continuous monitoring online experiments, leveraging real-time data to accelerate statistical inference during decision cycles.

A novel optimal design is developed for this unconventional paradigm, incorporating the equivalence theorem and structural characteristics.

In contrast to the typical sequential design which changes at each batch to achieve better exploration, we highlight the benefits of replication in continuous monitoring online experiments.

An explicit optimal design form for the classical online A/B testing problem is presented as a special case. Several numerical simulations validate the method's superiority, achieving marked enhancements in decision speed and estimation precision over traditional methodologies.

S052: High-Dimensional Inference and P-Value Combination

Jackknife Model Averaging with Quantile Regression under Distributed Heterogeneous Data

Haili Zhang

Shenzhen Polytechnic University

Quantile regression is important when covariate effects differ across outcome levels, but prediction under model uncertainty remains difficult. Model averaging provides a natural alternative to selecting a single specification. Existing jackknife model averaging methods, however, are developed for centrally stored data and are restrictive for distributed heterogeneous data, where raw observations cannot be pooled and local candidate submodels may differ across clients. We propose a distributed quantile regression jackknife model averaging (QR-JMA) procedure. Each client applies quantile partial correlation screening (QPCS) with a data-driven truncation rule to construct a sparse local candidate submodel, and weights are estimated by minimizing a global cross-validation criterion using FedProx. We establish asymptotic optimality in prediction risk for the infeasible exact weight under heterogeneous client distributions. For the finite-round FedProx weight, we derive a non-asymptotic bound on excess prediction risk with optimization error of order Op(Lk/T), where L is the number of clients, k the maximum candidate dimension, and T the number of communication rounds. Simulation studies and an Amazon review application show favorable predictive performance relative to competitors, reduced computation time relative to a centralized JMA benchmark, and the role of client-specific signal structure.

Statistical Inference of the Random Coefficients Multinomial Logit Model

• *Zixuan Wang, Jing Zhou, Junhao Hu, Daijun Wei, Qizhai Li*

South-Central Minzu University, Chinese Academy of Sciences, South-Central Minzu University, Hubei Minzu University, Chinese Academy of Sciences

The multinomial logit model is widely used for classification and discrete choice modeling, but is constrained by the Independence of Irrelevant Alternatives (IIA) assumption. The random coefficients multinomial logit model relaxes this restriction, offering greater flexibility in capturing heterogeneous individual preferences.

This talk investigates the application of the adaptive elastic-net to variable selection and parameter estimation in the random coefficients multinomial logit model, particularly under multicollinearity. We establish the Oracle properties of the resulting estimators, including asymptotic normality for nonzero parameters and consistent shrinkage of zero parameters. Estimation combines quasi-Monte Carlo simulation with coordinate descent optimization.

Simulation studies across multivariate normal, chi-squared, and t-distributed random coefficients show that the adaptive elastic-net outperforms both the adaptive LASSO and the standard elastic-net in accuracy and robustness. An empirical application to 2022 Chinese new energy SUV market data further validates the proposed approach, identifying price, vehicle class, powertrain type, and curb weight as key consumer preference drivers.

Feature Screening for Multiple Responses

• *Zhenzhen Jiang, Hongping Guo, Jinjuan Wang*

Hunan University of Technology and Business, Hubei Normal University, Beijing Institute of Technology

Feature screening has been widely investigated in many literatures and quite a few procedures have been proposed. However, most of the existing methods are developed based on regularization strategy and model assumptions such as linear model and Gaussian distribution, which limit their application range. And also, they were mainly designed to deal with univariate response and cannot handle multiple responses situations. To tackle these issues, we introduce a new association measure for multiple responses and univariate predictor, called multiple explained variability (MEV), and further propose a feature screening procedure, named MEV-SIS, based on MEV. MEV-SIS removes the commonly used model assumptions and can conduct feature screening for multiple responses simultaneously. The asymptotic properties of MEV are deduced, and the sure screening property and ranking consistency property of MEV-SIS are obtained. Extensive simulation studies and real data application demonstrate the advantage of MEV-SIS over the existing screening procedures in sufficiency and robustness.

S062: Machine Learning and Statistical Modeling Applications

TSPINN: Thompson Sampling-Based Adaptive Training for Physics-Informed Neural Networks

Dianpeng Wang

Beijing Institute of Technology

Physics-Informed Neural Networks (PINNs) offer a powerful framework for solving partial differential equations (PDEs) but often suffer from training inefficiencies. A critical factor is the selection of effective sampling distributions for training points. This work introduces Thompson Sampling PINN (TSPINN), a novel adaptive sampling approach inspired by Thompson

Sampling. TSPINN dynamically defines a reward distribution based on recent PDE residuals to guide the adaptive selection of training points, strategically balancing exploration of uncertain regions and exploitation of high-residual zones. Furthermore, we integrate TSPINN with the Causal PINN framework to develop Causal TSPINN, explicitly incorporating temporal and spatial causality into the sampling process. Extensive numerical experiments demonstrate that both TSPINN and Causal TSPINN significantly enhance solution accuracy while requiring fewer training iterations and collocation points compared to existing adaptive sampling techniques (e.g., R3, RAD). These methods demonstrate broad applicability across various PDE types, including Poisson, Helmholtz, and convection equations, among others.

Influence-Guided Subsampling Under Measurement Constraints

Lin Wang

Purdue University

Collecting labels for every point in a large dataset is often impractical due to measurement and budget constraints. A common remedy is to label only a carefully chosen subset of design points. In this talk, I will present a suite of computationally efficient subsampling methods that select small, informative subsets from large candidate pools. Unlike most existing approaches, which are tailored to low-dimensional settings, these methods explicitly accommodate high-dimensional regimes where the number of relevant predictors can be comparable to, or even exceed, the full sample size. I will describe theoretical guarantees and demonstrate performance through extensive numerical experiments. Across a wide range of scenarios, the methods consistently outperform existing subsampling techniques, resulting in substantial savings in labeling costs.

Mesh-Clustered Gaussian Process Emulator for Partial Differential Equation Systems

Chih-Li Sung, • Wenjia Wang, Liang Ding, Xingjian Wang

Michigan State University, National University of Singapore, Fudan University, Tsinghua University

Partial differential equations (PDEs) have become an essential tool for modeling complex physical systems. Such equations are typically solved numerically via mesh-based methods, such as the finite element method, the outputs of which consist of the solutions on a set of mesh nodes over the spatial domain. However, these simulations are often prohibitively costly to survey the input space. In this work, we propose an efficient emulator that simultaneously predicts the outputs on a set of mesh nodes, with theoretical justification of its uncertainty quantification. The novelty of the proposed method lies in the incorporation of the mesh node coordinates into the statistical model. In particular, the proposed method segments the mesh nodes into multiple clusters via a Dirichlet process prior and fits a Gaussian process model in each. Most importantly, by revealing the underlying clustering structures, the proposed method can extract valuable flow physics present in the systems that can be used to guide further investigations. Real examples are demonstrated to show that our proposed method has smaller prediction errors than its main competitors, with competitive computation time, and provides valuable insights about the underlying physics of the systems.

Noise Resistant Control Charts for Detecting Periodicity from Correlation

Yongxiang Li, Yunji Zhang, • Qian Xiao, Jianguo Wu

Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University, Peking University

Control charts are widely used to detect process variations that can reveal out-of-control events. Among various methods in the literature, spectral control charts are popular for monitoring processes with periodic behavior. Yet, in highly noisy environments, their performance may become unsatisfactory, because they are designed on the spectral features (e.g., periodogram), which could be easily masked by strong noise. Moreover, they would incorrectly report many correlated-but-not-periodic signals as periodic ones, leading to a high false discovery rate. To address these issues, we develop a new noise resistant feature in the time domain, based on which we propose two robust tests and establish noise resistant control charts for detecting periodicity from correlation. Our developed feature has been proved to have attractive finite-sample and asymptotic properties. In addition, we develop an efficient computation to calculate the proposed test statistics. Simulations and case studies are presented to demonstrate the superiority of the proposed approach compared to some state-of-the-art methods.

S063: Microbiome Intelligence: Large Models, Big Data, and Statistical Advances

BGOB: A Deep-Learning-Based Interpolation of Longitudinal Microbiome Data

• *Di Wu, Yixiang Qu, Ruiqi Lv*

University of North Carolina at Chapel Hill, University of North Carolina at Chapel Hill, Carnegie Mellon University

Longitudinal microbiome studies have been designed to study microbial community trajectories. The difficulty to analyze this types of data analysis may be due to their excess zeros, high noise,

and irregular sampling along time trajectory. These limitation make it challenging to perform many of the downstream analysis. Here, we introduce Bidirectional GRU-ODE-Bayes (BGOB), a deep learning framework developed for longitudinal microbiome interpolation. BGOB combines bidirectional information flow and ODE-based continuous modeling to jointly interpolate and smoothen trends across individual participants, providing uniform, denoised time intervals across patients. BGOB enables vastly improved performance in differential abundance testing and time-to-event analysis, and makes possible longitudinal analyses requiring uniformity, such as lead-lag detection and temporal clustering. After interpolation, previously low-powered datasets are able to broadly recapitulate known microbiology and elucidate interacting microbial communities. In simulations, we showed BGOB achieved high accuracy. In real data application, BGOB gives better power in association tests between microbial taxa and disease, and highlights novel species associated with Early Childhood Caries and disruption of key healthy gut microbiota in Inflammatory Bowel Disease.

Human Gut Mycobiome and Health in Large-Scale Cohort Studies

Ju-Sheng Zheng

Westlake University

We will describe our latest research progress in the gut mycobiome by using largest scale longitudinal cohort studies.

Integrative Statistical Analysis for Enhanced Scientific Discoveries in Metagenomics Studies

Xiang Zhan

Southeast University, School of Statistics and Data Science

The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing and shotgun metagenomic sequencing. Surprisingly, many microbiome studies have performed both experiments on the same cohort of samples. The two sequencing datasets often reveal consistent patterns of microbial signatures, suggesting that an integrative analysis of both datasets could enhance the testing power for these signatures. However, differential experimental biases, partially overlapping samples, and uneven library sizes pose tremendous challenges when combining the two datasets. In this article, we introduce the first method of this kind, named Com-2seq, that combines the two datasets for testing differential abundance at the genus level as well as the community level while overcoming these difficulties. Our simulation studies and real data application examples demonstrate that Com-2seq substantially enhances statistical efficiency over analysis of a single dataset and outperforms two ad hoc approaches to integrative analysis.

Application of Artificial Intelligence Content Generation (AIGC) in Big Data Mining of Microbiome

*Kang Ning, Yuan Wang

Huazhong University of Science and Technology, Huazhong University of Science and Technology

Microbial community is a kind of very complex form of life on earth, which participates in the material circulation and energy flow of the global ecosystem. Microbiome research has been the focus of scientists for a long time, and has made great progress in the past decade, but data acquisition, data management and data mining in the field of microbiome all remain challenging. With the rapid development of artificial intelligence (AI) technology, AIGC (Artificial Intelligence Generated Content) technology has shown great potential in data acquisition, data management and data mining. This work aims to explore the application of AIGC technology in the field of microbiome and provide new ideas for microbiome data acquisition, data management and data mining.

S069: Modern Statistical Inference and Modeling for Complex and Structured Data

A Study on the Performance of Matrix Control Charts Based on the Normal Distribution

Cong Li

Jilin University

With the explosive growth of data, matrix data has become increasingly common in fields such as genetic engineering, remote sensing satellites, and environmental and atmospheric sciences. Monitoring matrix data has also become increasingly important. Control charts are important tools in the study of statistical process control. Traditional research assumes that the monitored data follows a univariate normal distribution or a vector-valued normal distribution. Matrix sequences can explain the correlations between different row and column variables simultaneously. However, research on the monitoring of such sequences is very limited. In this article, based on the matrix normal distribution, we propose three new matrix control charts. We study the changes in the average run length of three control charts when a matrix normal process shifts. In the application of control charts, the true parameters are rarely known explicitly and are usually obtained through parameter estimation methods. Therefore, estimation errors are inevitable. This article further investigates the effect of parameter estimation errors on control charts. The application of the model and control charts is demonstrated in the example section.

Regression Analysis of Interval-Censored Failure Time Data

with Change Points and a Cured Subgroup

Yichen Lou, • Mingyue Du, Xinyuan Song

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

There exists a substantial body of literature that discusses regression analysis of interval-censored failure time data and also many methods have been proposed for handling the presence of a cured subgroup. However, only limited research exists on the problems incorporating change points, with or without a cured subgroup, which can occur in various contexts such as clinical trials where disease risks may shift dramatically when certain biological indicators exceed specific thresholds. To fill this gap, we consider a class of partly linear transformation models within the mixture cure model framework and propose a sieve maximum likelihood estimation approach using Bernstein polynomials and piecewise linear functions for inference. Additionally, we provide a data-driven adaptive procedure to identify the number and locations of change points and establish the asymptotic properties of the proposed method. Extensive simulation studies demonstrate the effectiveness and practical utility of the proposed methods, which are applied to the real data from a breast cancer study that motivated this work.

Quantile Regression for Censored Data with a Cure Fraction

Bo Han

Yunnan University

Quantile regression is emerging as a valuable statistical tool for exploring complete information on the exposure-outcome relationship. This paper investigates a novel inference procedure for quantile regression models to analyze survival data with a cure fraction. A twocomponent mixture cure model is used, in which the cure fraction follows a logistic model and the uncured individuals follow a linear quantile regression model. Our main interest lies in the estimation of regression coefficients in the latter model. This is achieved by minimizing a loss function that has three building blocks: (1) the check function, commonly used in quantile regression, (2) a first weight function that redistributes the weight of censored observations, and (3) a second weight function that accounts for the latent cure status. The minimization of this loss function leads to a computationally simple estimation procedure for quantile regression parameters. The approach of dimension reduction is employed to circumvent the curse of dimensionality. Asymptotic properties of the resulting estimators are established, including the consistency and weak convergence. Simulation studies, including homogeneous and heterogeneous error cases, demonstrate the advantageous empirical performance of the proposed method. The analysis of a breast cancer study illustrates the practical utility of our cure rate quantile regression.

S070: Modern Statistical Learning and High Dimensional Modeling

Acceleration for Diffusion Models: Theory and Practice

Gen Li

The Chinese University of Hong Kong

Diffusion models, which generate new data instances by learning to reverse a Markov diffusion process from noise, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain underdeveloped. Particularly, despite the recent surge of interest in accelerating sampling speed, convergence theory for these acceleration techniques remains limited. In this talk, I will first introduce an acceleration sampling scheme that provably improves the iteration complexity under minimal assumptions. I will then present a new first-order solver for diffusion sampling that achieves performance comparable to higher-order methods, offering a fundamentally new pathway for accelerating diffusion model sampling.

High-Dimensional Longitudinal Model Averaging Via Leave-Subject-Out Cross-Validation Functionals

Shirong Deng

Wuhan University

Model averaging is an effective way to enhance prediction accuracy. However, most existing works focus on low-dimensional settings with cross-sectional data, and its application to high-dimensional longitudinal data remains very limited. To this end, we propose three novel model averaging methods for high-dimensional longitudinal data. Specifically, considering the semiparametric model with splitting covariates as candidate models, we define the leave-subject-out cross-validation (LsoCV) process and further propose three novel LsoCV functionals, including the end-time LsoCV, integrated LsoCV, and supremum LsoCV, to achieve more accurate prediction for longitudinal response. The optimal weights for combining candidate models can be obtained by minimizing any of these functionals, and the resulting highdimensional optimization problem is solved by a greedy algorithm. Theoretically, we show that when all candidate models are misspecified, all three proposed model averaging methods can attain the lowest possible prediction loss asymptotically. When correctly specified models are included among the candidates, we prove that the summation of weights assigned to these correctly specified models converges to one in probability. Finally, extensive simulation studies and an application to a microbiome dataset demonstrate that the proposed methods achieve superior prediction accuracy over existing methods.

A Conditional Distribution Equality Testing Framework Using Deep Generative Learning

Siming Zheng, ⋆ *Tong Wang, Meifang Lan, Yuanyuan Lin*

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

We propose a general framework for testing the conditional distribution equality in a two-sample problem, which is most relevant to covariate shift and causal discovery. Our framework is built on neural network-based generative methods and sample splitting techniques by transforming the conditional testing problem into an unconditional one. We introduce the generative classification accuracy-based conditional distribution equality test (GCA-CDET) to illustrate the proposed framework. We establish the convergence rate for the learned generator by deriving new results related to the recently-developed offset Rademacher complexity and prove the testing consistency of GCA-CDET under mild conditions. Empirically, we conduct numerical studies including synthetic datasets and two real-world datasets, demonstrating the effectiveness of our approach. Additional discussions on the optimality of the proposed framework are provided in the online supplementary material.

Deep Nonparametric Inference for Conditional Mean Models with Panel Count Data

⋆ *Anyin Feng, Qiang Wu, Xiangbin Hu and Xingqiu Zhao*

The Hong Kong Polytechnic University

In the statistical analysis of panel count data, semiparametric models such as the proportional mean model and the accelerated mean model are often used to investigate the occurrence rate of recurrent events. Misspecification of the prespecified model may lead to biased estimation and poor predictive performance. Therefore, it is essential to develop hypothesis testing methods to assess the adequacy of the assumed model. However, most existing diagnostic methods rely on graphical tools and lack theoretical justification. To address this issue, we develop a hypothesis testing method to assess model adequacy with theoretical support. Specifically, we first propose a deep extended mean model that integrates commonly used nonparametric and semiparametric models within a general framework. Treating the proposed model as the encompassing model, the hypothesis test is constructed based on the discrepancy between the parameter estimates obtained under the encompassing and prespecified models. For the deep extended mean model, we derive a nonasymptotic error bound for the least-squares-based estimator and prove the asymptotic normality of functionals of the estimator. Building on these results, we establish the asymptotic normality of the proposed test statistic under the null hypothesis, thereby providing a theoretical foundation for the testing procedure. Extensive simulation studies and an application to data from a skin cancer chemoprevention trial demonstrate the superior performance of the proposed method.

S078: New Advances in Differential Privacy

Truth Inflated Mechanism for Distributional Privacy

Yanqing Zhang

Yunnan University

Differential privacy provides a rigorous framework for individual data protection but often suffers from degraded utility and practical limitations in real-word statistical analysis.

Its reliance on neighboring dataset, sensitivity bounds and independent noise makes it ill-suited for complex datasets or correlated structure.

To address these challenges, the paper introduces a distribution-aware privacy definition, called as distributional privacy, which guarantees a randomized mechanism's outputs are statistically indistinguishable for any two datasets drawn from the same underlying distribution.

This approach handles correlated data and complex data structures by focusing on distributional similarly rather than individual record differences.

Moreover, the paper propose the Truth Inflated Mechanism (TIM) and the specific algorithm as a concrete method to achieve ϵ-distributional privacy by probabilistically releasing the true query from a finite set based on the empirical query distribution, with probabilities strategically inflated towards the true query result.

Theoretical analysis establishes that TIM satisfies ϵ-distributional privacy and provides an upper bound on the utility of the query that decays exponentially with the privacy budget.

Our numerical studies also demonstrate superb performance of the proposed method on the utility guarantee of the privacy-preserving data released.

Online Differentially Private Inference for Linear Regression Model

⋆ *Senlin Yuan, Fang Liu, Xuerong Chen*

Yunnan University

In the era of big data, data privacy has attracted increasing attention. Differential privacy (DP) is a state-of-the-art framework for formal privacy guarantees. Many privacy-preserving inference methods have been developed for releasing information from a wide range of data analyses in the DP framework. However, DP statistical inference methods for streaming data, which represent a common type of big data, are still lacking. In this paper, we propose a computationally efficient privacy-preserving method for online update and inference of linear regression models that satisfies DP guarantees. We derive regression parameter estimates

in the DP framework, along with the covariance estimates based on which privacy-preserving confidence intervals for the parameters are constructed. We provide theoretical support for the proposed DP method, and numerical results demonstrate the good performance of our approach.

Data Aggregation Based on Adaptive Personalized Local Differential Privacy

Peng Zhang

Kunming University of Science and Technology

In practical scenarios, the absence of a trusted center poses a significant challenge to collecting data while preserving data privacy. In this paper, we propose an Adaptive Personalized Local Differential Privacy (APLDP) method for estimating the frequency of different values in discrete data and the mean of continuous data. Our approach allows each individual to have a different privacy budget, and these budgets are related to covariates. Notably, the existing personalized differential privacy framework emerges as a special case of our proposed method. We demonstrate that both the frequency and mean estimates obtained via APLDP are unbiased and satisfy (ε,δ)-LDP. Furthermore, we provide theoretical analysis that establishes the statistical utility of the estimates. Finally, we validate the utility of the proposed method through numerical simulations and real data analysis.

A General Distributed Privacy-Preserving Statistical Inference Framework for Massive Data

Senlin Yuan, ⋆ *Xuerong Chen, Yong Zhou*

Yunnan University, Southwestern University of Finance and Economics, East China Normal University

In the era of massive data, privacy-preserving distributed inference plays a key role in balancing data utility, computational efficiency, and privacy protection in statistical analysis.

However, current differentially private distributed inference and federated learning methods either focus solely on parameter estimation, assume i.i.d. data, or are restricted to specific models (e.g., linear models), failing to provide comprehensive statistical inference for general models with heterogeneous data. Moreover, sensitivity calculations in existing differential privacy methods often rely on data truncation or boundedness assumptions, tailored to each model on a case-by-case basis. To address these issues, we propose a unified distributed privacy-preserving statistical inference framework suitable for both homogeneous and heterogeneous data. The framework leverages an encryption technique that integrates concepts from symmetric encryption and differential privacy. It is computationally efficient, achieves high accuracy, and eliminates the need for case-by-case sensitivity calculations. The framework is applicable to widely used models such as linear regression, logistic regression, the Cox model, and quantile regression, among others.

We also provide guidance on selecting encryption functions that can simplify both sensitivity calculations and theoretical derivations. Theoretical properties, including differential privacy guarantees, convergence rates, and the asymptotic normality of the proposed private estimators, are established. The effectiveness of our method is demonstrated through both numerical simulations and real data analyses.

S082: New Developments of Statistical Methods in Bio-Medical Studies

Model Selection for Archimedean Copula Models Copula Models Based on Dependent Censored Data

Antai Wang

New Jersey Institute of Technology

In this talk, we first present the identifiability property for any models belonging to Archimedean copula model familybased on dependent competing risks data when at least one of the failure times is exponentially distributed. The result isa generalization of the identifiability property proved in Wang (2023). Using this property, we propose a model selectionprocedure for Archimedean copula models based on dependent censored data. We demonstrate the proposed procedureusing simulation studies and in an application to survival data.

Adaptive Metric Learning on SPD Manifolds via the Alpha–Procrustes Framework

⋆ *Shao-Wei Huang, Jingyi Zheng, Huajun Huang*

Auburn University, Auburn University, Auburn University

In neuroimaging analysis, correlation matrices are widely used to characterize functional connectivity. These symmetric positive definite (SPD) matrices lie on a Riemannian manifold, where the choice of metric induces the underlying geometry. Commonly used metrics, such as the Bures–Wasserstein (BW) and Log-Euclidean metrics, provide distinct yet fixed geometries that may not fully capture the intrinsic variability of functional connectomes. To address this limitation, we propose a generalized metric-learning pipeline based on the Alpha-Procrustes (AP) metric. By introducing a parameter α, the AP metric spans a family of geometries, with BW and Log-Euclidean metrics as special cases. This enables data-driven selection of an optimal geometry tailored to the underlying structure of the data. Under the learned metric, SPD matrices are projected onto the corresponding tangent space, facilitating the application of classical machine learning models. We evaluate the proposed

framework on both brain–computer interface (BCI) and fMRI datasets. Experimental results show that the optimal choice of α varies across subjects, and that variations in the induced geometry lead to substantial differences in predictive performance. These findings underscore the importance of adaptive metric selection and suggest that AP-based frameworks provide an effective strategy for improving generalization and mitigating domain shift in neuroimaging applications.

Interval Estimation for the Area Under the ROC Curve with Verification Bias

Shirui Wang, Shuangfei Shi, •Gengsheng Qin

Georgia State University, Georgia State University, Georgia State University

In medical diagnostic studies, the area under the receiver operating characteristic curve (AUC) is a widely used metric that captures a continuous test's overall ability to discriminate between diseased and non-diseased individuals across all possible cutoffs. However, in practice, disease status is sometimes only partially verified, introducing verification bias that undermines the validity of AUC estimation. While numerous methods address bias correction for AUC estimation, approaches that directly construct confidence intervals for the AUC remain limited. This paper proposes two robust methods for constructing bias-corrected confidence intervals for the AUC under the missing-at-random (MAR) assumption: one based on bootstrap resampling and the other on empirical likelihood. Both approaches accommodate missing disease verification by leveraging the bias-corrected ROC estimators introduced by Alonzo and Pepe. Extensive simulation studies and real-world data analyses demonstrate that our proposed methods yield valid and precise interval estimates for the AUC under various clinically relevant settings.

Latent Single-Index Models with Factor Structure for Multivariate Ordinal Data

Zhiyong Chen

Fujian Normal University

We propose a latent semi-parametric model for multivariate ordinal data, in which the single-index models with factor structure are used to assess the effects of the latent covariates on the latent responses and to explore the covariance structure of the latent responses. Since the indices may vary from minus infinity to plus infinity, the traditional spline can not be applied directly to approximate the unknown link function. We employ a modified version to tackle this pervasive problem by transforming the indices into the unit interval via a continuously cumulative distribution function and constructing a novel combination of the spline. A fully Bayesian modelling is performed by facilitating efficient MCMC approach with free-knot splines to analyze the proposed model. We also consider a proposed distribution augmentation algorithm to sample from the conditional posterior of the correlation matrix. To accelerate the convergence, we make use of the partial-collapse and parameter expansion as well as reparameterization techniques in our algorithm. The performance of the proposed model and estimation method are checked through a simulation study and applied to a real dataset.

S086: Novel Statistical and Machine Learning Methods for Analyzing Omics Data

Contextual Evaluation of MicroRNA Sequencing Data Harmonization in Machine Learning Tasks

Li-Xuan Qin

Memorial Sloan Kettering Cancer Center

Translating microRNA sequencing into reliable sample classifiers requires appropriate data harmonization to counter artifacts arising from heterogeneous experimental handling. To inform practical use, we systematically evaluated harmonization methods for machine-learning-based sample classification , using a framework that integrates AI-augmented datasets, statistical evaluation pipelines, and accessible software tools. Across a range of signal-to-artifact ratios, harmonization often partially recovered accuracy lost to artifacts, with gains that varied by machine learning approach. Simple scaling methods, notably TMM and Total Count, performed consistently well, whereas RUV variants generally underperformed. We also found that applying 'frozen' quantile normalization to the validation set, using transformation learned from the training set, outperformed its standard form. Reanalysis of breast cancer data from The Cancer Genome Atlas corroborated these findings. Overall, this study underscores the need for context-specific harmonization in machine learning classification workflows and provides concrete guidance on effective harmonization-classifier pairings to advance predictive genomics for precision medicine.

Exploration of Innovative Deep Learning Methods for Cancer Studies Based on Multi-Omics Data

•*Zheng Xu, Cong Wu, Yan Su, Changzhi Ma*

Wright State University, Wright State University, Medpace, College of Wooster

Cancer is a complex and heterogeneous disease related to multiple factors, including genomics, transcriptomics, epigenomics, proteomics and other variables. Cancer analysis often involves multiple complex bioinformatics datasets for better discovery of mechanisms, predicts cancer survival, and other objectives such as cancer stage classification. Machine learning and deep

learning have seen rapid progress in cancer analysis. We aimed to explore some innovative deep learning methods to evaluate their performance. The performance of different deep learning methods will be evaluated on publicly available cancer datasets especially Breast Cancer data from the Cancer Genome Atlas based on the tasks including cancer classification (e.g. subtype and stage) and survival prediction.

A Divide and Conquer Strategy for Recapitulating Whole Genome 3D Structure Using Hi-C Data

Jincheol Park, ⋆Meng Wang, Emerson Webb, Victor Jin, Shili Lin

Abigail Wexner Research Institute at Nationwide Children's Hospital

The three dimensional (3D) spatial organization of the genome is closely linked to biological functions and can be captured by Hi-C assays through interrogating genome-wide chromatin interactions. Methodologies for inferring 3D structures from Hi-C data summarized as a two-dimensional (2D) contact matrix can be broadly placed within the paradigms of optimization-based and sampling-based. Many optimization-based methods are capable of constructing whole genome 3D structures but do not account for spatial dependency in the 2D data matrix nor cell heterogeneity in bulk Hi-C data, which provide an average over millions of cells. Sampling-based methods, on the other hand, are probabilistic model-based and can account for not only dependency, heterogeneity, but also other features inherent in Hi-C data, such as over-dispersion and sparsity. However, whole-genome 3D structure recapitulation is too computationally expensive for sampling-based methods, while chromosome-by-chromosome strategies for sampling-based methods ignore important information on inter-chromosomal contacts. To address these issues, we propose the truncated Random effect EXpression-cut and paste (tREX-cap) method, which applies the tREX model within a divide and conquer strategy. The resulting method inherits the good data-feature-cognizant properties of tREX and, in the meantime, can efficiently infer the whole genome 3D structure. We demonstrate the performance of tREX-cap through an extensive simulation study and analysis of a Hi-C lymphoblastoid dataset.

Bayesian Methods for Phenotype-Associated Domain and Gene Detection in Spatial Transcriptomics

Han Wang

China Agricultural University

Spatial transcriptomics enables the simultaneous profiling of gene expression and tissue architecture, yet linking localized spatial domains to macro-scale clinical phenotypes and identifying spatially differential genes remain statistically challenging. Here we present two Bayesian frameworks addressing these tasks. First, we introduce SPhenX, a spatially constrained matrix factorization approach that embeds bulk transcriptomes onto spatial domain coordinates through Bayesian methods, bridging the modality gap between large phenotyped cohorts and spatially resolved references to quantify per-domain contributions to clinical outcomes without requiring cell-type annotation. Second, we develop a hierarchical Bayesian spatial latent factor model that decomposes multi-slice spatial transcriptomics data into group-specific and shared latent components, enabling the identification of differential spatial domains and their driving gene modules between disease and control conditions via scalable variational inference. Together, these methods demonstrate that principled Bayesian modeling of spatial dependencies—whether through CAR-regularized cross-modal integration or GP-based latent factor decomposition—substantially improves the detection of clinically relevant tissue architectures and spatially resolved molecular programs across diverse biological contexts.

S099: Recent Advances in High-Dimensional and Functional Hypothesis Testing

A Unified Estimation and Inference Framework for High-Dimensional Time Series Models

Jinyuan Chang, ⋆Qiao Hu, Zhentao Shi, Jia Zhang

Southwestern University of Finance and Economics, The Chinese University of Hong Kong, Chinese University of Hong Kong, Southwestern University of Finance and Economics

Conomic and financial models, such as vector autoregressions, local projections, and multivariate volatility models, involve intricate dynamic interactions and spillovers across numerous time series. These models can be integrated into a unified framework, with high-dimensional parameters identified by associated moment conditions. As the number of parameters and moment conditions may surpass the sample size, we pro- pose adding a double penalty to the empirical likelihood criterion to induce sparsity and facilitate dimension reduction. Notably, we utilize a marginal empirical likelihood approach even when data exhibits temporal dependence. Under regularity conditions, we provide asymptotic guarantees for our method, making it an attractive option for estimating large-scale multivariate time series models. We demonstrate the versatility of our procedure through extensive Monte Carlo simulations and three empirical appli- cations, including an analysis of the US sectoral inflation rates, fiscal multipliers, and volatility spillover across China's banking industry.

High-Dimensional Convoluted Rank Regression with Latent Factors

•*Chenye Zhao, Yanmei Shi, Xu Guo*

Beijing Normal University, Beijing Normal University, Beijing Normal University

In recent years, high-dimensional linear regression models with latent factors have garnered significant attention in the statistical literature. However, existing methods often perform poorly when the data contain outliers or exhibit heavy-tailed distributions. To address these challenges, we propose a convoluted rank regression estimator that is effective under significant dependence among covariates and remains robust in the presence of outliers or heavy-tailed errors. We establish non-asymptotic error bounds under mild regularity conditions. Extensive simulations demonstrate the favorable finite-sample performance of the proposed method, and an application to the FRED-MD dataset highlights its robustness and practical relevance.

Goodness-of-Fit Test for Generalized Functional Linear Models via Projection Averaging

Feifei Chen, Kaiming Zhang, • Yanni Zhang, Hua Liang

Beijing Normal University, Beijing Normal University, Beijing Normal University, George Washington University

Assessing model adequacy is a crucial step in regression analysis, ensuring the validity of statistical inferences. For Generalized Functional Linear Models (GFLMs), which are widely used for modeling relationships between scalar responses and functional predictors, there is a recognized need for formal goodness-of-fit testing procedures. Current literature on this specific topic remains limited.

This paper introduces a novel goodness-of-fit test for GFLMs. The test statistic is formulated as a U-statistic derived from a Cramér-von-Mises metric integrated over all one-dimensional projections of the functional predictor. This projection averaging strategy is designed to effectively mitigate the curse of dimensionality. We establish the asymptotic normality of the test statistic under the null hypothesis and prove the consistency under the alternatives. As the asymptotic variance of the limiting null distribution can be complex for practical use, we also propose practical bootstrap resampling methods for both continuous and discrete responses to compute p-values. Simulation studies confirm that the proposed test demonstrates good power performance across various settings, showing advantages over existing methods.

S100: Recent Advances in High-Dimensional Data Analysis

Efficient Machine Unlearning With Minimax Optimality

Jingyi Xie, Linjun Zhang, •Sai Li

Renmin University of China, Rutgers University, Tsinghua University

There is a growing demand for efficient data removal to comply with regulations like the GDPR and to mitigate the influence of biased or corrupted data. This has motivated the field of machine unlearning, which aims to eliminate the influence of specific data subsets without the cost of full retraining. In this work, we propose a statistical framework for machine unlearning with generic 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 forget model bias. 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.

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

Shiyang Ma

Shanghai Jiao Tong University

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.

Experimental Design and Differential Inference for Comparative Single-Cell RNA-Sequencing Studies

•*Fangda Song, Kevin Y. Yip, Yingying Wei*

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

Single-cell RNA-sequencing (scRNA-seq) experiments are becoming increasingly complicated with multiple treatments or biological conditions. However, guidelines on experimental designs and rigorous statistical methods for comparative scRNA-seq studies with cells collected from multiple conditions are still lacking. For a confounded design, the batch effects, cell-type effects, and condition effects can never be distinguished. Therefore, we mathematically derive the requirements for a valid comparative scRNA-seq experimental design. Moreover, existing methods for identifying differentially expressed genes and differential cell-type abundances between conditions have to be multi-stage. Because multi-stage approaches ignore uncertainties in previous stages and may propagate errors from earlier stages to later stages, they can suffer from high error rates. Here, we introduce DIFseq, a hierarchical model that accounts for all uncertainties and hence rigorously quantifies the condition effects on both cellular composition and cell-type-specific gene expression levels. DIFseq substantially outperforms state-of-the-art methods for both simulated and real data.

Information-Incorporated Estimation of High-Dimensional Generalized Linear DAGs with Interventions

•*Weijuan Liang, Qingzhao Zhang, Shuangge Ma*

Xiamen University, Xiamen University, Yale School of Public Health

We study the problem of regularized estimation for the structure and parameters of high-dimensional sparse generalized linear directed acyclic graphs (GLDAGs). Theoretically, we demonstrate that estimation based solely on observational data can only recover the observational Markov equivalence class, leaving a large portion of the DAG unidentifiable. To address this limitation, we leverage both observational and interventional data to develop an estimation method for the interventional Markov equivalence class of GLDAGs, which achieves finer identifiability due to the additional information from interventions. Furthermore, we incorporate external information from prior studies to enhance estimation in the current data. We establish that informative priors not only improve the convergence rate of estimation but also facilitate the identification of weaker signals. In addition to methodological development and theoretical guarantees, we provide empirical results to demonstrate the effectiveness of our approach.

S115: Recent Advances on the Analysis of Failure Time Data

Longitudinal First Hitting-Time Models for Survival Data

Mei-Ling Ting Lee

University of Maryland

First-hitting time (FHT) threshold regression (TR) is well-known for analyzing event time data without the proportional hazards assumption. To date, most applications and software are developed for cross-sectional data. In this paper, using the Markov property of processes with stationary independent increments, we present methods and procedures for conducting longitudinal threshold regression (LTR) for event time data with or without covariates. We demonstrate the usage of LTR in two case scenarios, namely, analyzing laser reliability data without covariates, and cardiovascular health data with time-dependent covariates. Moreover, we provide a simple-to-use R function for LTR estimation for applications using Wiener processes.

Beyond Hazard-Based Illness-Death Models for Semi-Competing Risks Data

Rui Huang, •*Liming Xiang*

Nanyang Technological University, Nanyang Technological University

The illness-death model has been developed as a common approach for regression analysis of semicompeting risks data by modeling transitions between states using the proportional hazards or accelerated failure time model. In this work, we propose an illness-death model based on the mean residual life function, which naturally captures the state-dependent structure of the semicompeting risks process by conditioning on disease progression having occurred and provides a clinically meaningful and interpretable framework for modeling state transitions. We facilitate estimation through novel estimating equations derived from a penalized quasi-likelihood approach incorporating inverse probability weighting. Our method requires no distributional assumption on the latent frailty term, thereby reducing the risk of model misspecification. Simulation studies demonstrate its promising performance in across a range of realistic scenarios and an application to a myeloma progression study illustrates its practical utility.

Prognostic Assessment of High Dimensional Survival Targets with Auxiliary Information Transfer

Jie Ding, Jialiang Li, •*Xiaoguang Wang*

Dalian Maritime University, National University of Singapore, Dalian University of Technology

The adaptive utilization of external informative sources is

becoming increasingly crucial for enhancing the statistical analysis of an internal target. This paper presents a novel information synthesis framework capable of transferring aggregate statistics in high dimensional survival cohorts, thereby improving the assessment of prognostic factors. Most of the existing literature requires sharing individual-level records from the outside, while this is not necessary for our methodology, resulting in strong privacy protection. The control variate technique is employed to obtain estimators of interested effects, and its essential validity condition is guaranteed through double decorrelation-aided adjustments. To ensure computational stability, a refinement based on sample splitting is further carried out. Asymptotically, our estimator is more efficient than traditional estimators that rely solely on the target. When the dimension is low and fixed, it degenerates to one that has exactly the same asymptotic efficiency as the well-known empirical likelihood-based estimator, yet it maintains a significantly lower computational complexity. We demonstrate the improved performances of our method in terms of estimation efficiency and testing power through extensive simulations and a real-world case study.

An Optimal Two-Step Estimation Approach for Two-Phase Studies

•*Qingning Zhou, Kin Yau Wong*

The University of North Carolina at Charlotte, The Hong Kong Polytechnic University

Two-phase sampling is commonly adopted for reducing cost and improving estimation efficiency. In many two-phase studies, the outcome and some cheap covariates are observed for a large sample in Phase I, and expensive covariates are obtained for a selected subset of the sample in Phase II. As a result, the analysis of the association between the outcome and covariates faces a missing data problem. Complete-case analysis, which relies solely on the Phase II sample, is generally inefficient. In this paper, we study a two-step estimation approach, which first obtains an estimator using the complete data, and then updates it using an asymptotically mean-zero estimator obtained from a working model between the outcome and cheap covariates using the full data. This two-step estimator is asymptotically at least as efficient as the complete-data estimator and is robust to misspecification of the working model. We propose a kernel-based method to construct a two-step estimator that achieves optimal efficiency. Additionally, we develop a simple joint update approach based on multiple working models to approximate the optimal estimator when a fully nonparametric kernel approach is infeasible. We illustrate the proposed methods with various outcome models. We demonstrate their advantages over existing approaches through simulation studies and provide an application to a major cancer genomics study.

S121: Scalable and Distributed Statistical Learning for Complex Data Structures

Synthetic Data for Imbalanced Learning

•*Anru Zhang, Zhengchi Ma*

Duke University, Duke University

Imbalanced data arise widely in modern applications, where one class is much rarer than the other and standard learning methods often favor the majority class at the expense of rare but important outcomes. A widely used strategy is to generate synthetic minority samples and train prediction models on the augmented data. Although effective in practice, the statistical benefits of this approach remain poorly understood: synthetic augmentation may reduce the effects of imbalance, but it may also introduce bias if the generated samples do not accurately reflect the target distribution.

In this talk, I will present a statistical perspective on synthetic augmentation for imbalanced learning. I will describe a unified theoretical framework that clarifies when synthetic minority samples improve performance, when they can be harmful, and how the optimal amount of augmentation depends on generator accuracy and the structure of the learning problem. Our results show that the benefit of augmentation is governed by the interaction between class asymmetry, synthetic bias, and the underlying prediction task. These insights also lead to a practical, data-driven method for choosing the synthetic sample size. I will conclude with several open questions and directions for future work.

Covariance Graph-Guided Harmonization in Distributed Neuroimaging Studies

Yiwen Dong, Jingsong Zhou, Shuo Chen, Dana Tudorascu, •*Qiong Wu*

University of Pittsburgh, University of Maryland, University of Maryland, University of Pittsburgh, University of Pittsburgh

Multi-site neuroimaging studies enhance statistical power and generalizability but face challenges from data privacy constraints where direct data sharing is prohibited. In addition, neuroimaging data across multiple institutions often exhibit systematic heterogeneity driven by scanner differences, which could bias downstream analyses if not addressed. We propose a distributed covariance graph-guided ComBat (dG-ComBat), a multivariate harmonization framework that operates in a fully privacy-preserving distributed setting. dG-ComBat identifies and leverages the latent covariance graph structure among neuroimaging features to guide mitigation of both mean and covariance batch effects. With only two rounds of summary-statistics exchange, dG-ComBat achieves identical harmonization performance

(i.e., lossless) to the centralized setting in which data from all participating institutions are shared. Extensive simulations and applications to the Alzheimer's Disease Neuroimaging Initiative (ADNI) showed that dG-ComBat effectively reduces scanner-related variability while preserving biologically relevant diagnostic signals and improving downstream analyses.

Bayesian Intelligence for Brain–Computer Interfaces

*Jian Kang

University of Michigan

Brain–computer interfaces (BCIs) enable direct communication by decoding neural signals, providing important assistive technology for individuals with severe motor impairments. From a statistical perspective, BCIs present several challenges, including low signal-to-noise ratios, complex spatial–temporal dependence in neural recordings, substantial heterogeneity across users, and the need to balance prediction accuracy with communication speed. In this talk, I will present a set of Bayesian modeling and learning approaches for EEG P300-based BCIs that aim to address these challenges in a unified probabilistic framework. These include probabilistic models for neural responses, Bayesian classification methods that exploit latent structure and sparsity for identifying informative neural features, adaptive decision strategies for sequential stimulus selection and early stopping, and transfer learning approaches that leverage information across users to reduce calibration time. These developments illustrate how Bayesian modeling, uncertainty quantification, and adaptive decision-making can provide a principled foundation for building more accurate, interpretable, and efficient BCIs.

S135: Statistical Innovations for Spatial Transcriptomics and 3D Genomics

When Sparsity Meets Spatial Structure: Fusion of Sparsity and Spatial Information for Clustering Hi-C Single-Cell Data

Shili Lin

Ohio State University

Clustering is a fundamental tool across many scientific disciplines, particularly in genomics. However, traditional clustering methods such as K-means often struggle with the curse of dimensionality inherent in high-throughput genomic data. Single-cell genomic data introduce additional challenges characterized by Sparsity and Spatial structure (S^2).

In this talk, we present methods that leverage these unique features to enhance clustering performance. Specifically, we propose an extension of Kendall's tau distance that integrates S^2 information in the context of single-cell Hi-C data. Furthermore, we explore strategies for early and late integration of S^2 features using the Jensen–Shannon divergence measure. Together, these approaches demonstrate how incorporating sparsity and spatial structure can yield more biologically meaningful clustering results in complex genomic datasets.

Unveiling Fine-Scale Spatial Structures and Amplifying Gene Expression Signals in Ultra-Large ST Slices with HERGAST

Yuqiao Gong, Xin Yuan, Zhangsheng Yu

Shanghai Jiao Tong University, Shanghai Jiao Tong University, Shanghai Jiao Tong University

We propose HERGAST, a system for spatial structure identification and signal amplification in ultra-large-scale and ultra-high-resolution spatial transcriptomics data. To handle ultra-large spatial transcriptomics (ST) data, we consider the divide and conquer strategy and devise a Divide-Iterate-Conquer framework especially for spatial transcriptomics data analysis, which can also be adopted by other computational methods for extending to ultra-large-scale ST data analysis. To tackle the potential over-smoothing problem arising from data splitting, we construct a heterogeneous graph network to incorporate both local and global spatial relationships. In simulations, HERGAST consistently outperforms other methods across all settings with more than a 10% increase in average adjusted rand index (ARI). In real-world datasets, HERGAST's high-precision spatial clustering identifies SPP1+ macrophages intermingled within colorectal tumors, while the enhanced gene expression signals reveal unique spatial expression patterns of key genes in breast cancer.

Statistical Methods for Cell-Cell Interaction Studies on Spatially Resolved Transcriptomics

Xiaoyu Song

National University of Singapore

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.

Spatially Aware Adjusted Rand Index for Evaluating Spatial Transcriptomics Clustering

Yinqiao Yan, Xiangnan Feng, ⋆Xiangyu Luo

Beijing University of Technology, Fudan University, Renmin University of China

The spatial transcriptomics (ST) clustering plays a crucial role in elucidating the tissue spatial heterogeneity. An accurate ST clustering result can greatly benefit downstream biological analyses. As various ST clustering approaches are proposed in recent years, comparing their clustering accuracy becomes important in benchmarking studies. However, the widely used metric, adjusted Rand index (ARI), totally ignores the spatial information in ST data, which prevents ARI from fully evaluating spatial ST clustering methods. We propose a spatially aware Rand index (spRI) as well as spARI that incorporate the spatial distance information. Specifically, when comparing two partitions, spRI provides a disagreement object pair with a weight relying on the distance of the two objects, whereas Rand index assigns a zero weight to it. This spatially aware feature of spRI adaptively differentiates disagreement object pairs based on their distinct distances, providing a useful evaluation metric that favors spatial coherence of clustering. The spARI is obtained by adjusting spRI for random chances such that its expectation takes zero under an appropriate null model. Statistical properties of spRI and spARI are discussed. The applications to simulation study and two ST datasets demonstrate the improved utilities of spARI compared to ARI in evaluating ST clustering methods.

S139: Statistical Learning with AI

A Sparse Bayesian Learning Algorithm for Estimation of Interaction Kernels in Motsch-Tadmor Model

⋆*Jinchao Feng, Sui Tang*

Great Bay University, University of California, Santa Barbara

In this talk, I will present a data-driven framework for identifying asymmetric interaction kernels in the Motsch–Tadmor model based on observed agent trajectories. Unlike symmetric models, the asymmetric setting introduces a nonlinear inverse problem due to the normalization of interaction weights. We reformulate the problem using the implicit form of the governing equations, reducing kernel learning to a subspace identification task. To solve this, we develop a sparse Bayesian learning approach that incorporates prior structure and quantifies uncertainty, enabling robust model selection under noise. Numerical experiments on several prototype systems demonstrate the method's ability to recover key interaction patterns and predict collective behavior accurately, even with limited or noisy data.

Score-Based Generative Models with Gamma Processes

Libai Xu

Soochow University

Diffusion models are predominantly built on Gaussian perturbations, whose symmetry and light-tailed behavior provide substantial analytical and algorithmic convenience. However, this design choice can be restrictive when the target perturbation structure is asymmetric, skewed, or exhibits localized abrupt variations. We present Gamma-Noise Diffusion Models (GNDM), a non-Gaussian diffusion framework that replaces Gaussian perturbations with gamma-driven increments while preserving the familiar score-based training and reverse-time sampling interface. Our approach constructs a forward corruption process with composable gamma perturbations, derives an approximate reverse-time stochastic dynamics using a local expansion of the reverse conditional, and further develops a probability-flow ordinary differential equation (ODE) that enables deterministic sampling. We also introduce a weighted score-matching objective adapted to the conditional gamma perturbation family. Experiments on MNIST, CIFAR-10, and CelebA demonstrate that the proposed framework is practically trainable and capable of high-quality generation. On unconditional CIFAR-10 generation, our implementation achieves an FID of 2.58 and an IS of 10.02. These results suggest that diffusion modeling can be extended beyond Gaussian assumptions without discarding the core score-based generative modeling pipeline.

Neural Dynamic Time Warping A Differentiable Framework for Financial Bubble Analysis

⋆*Lukun Zheng, Yufei Zheng*

Western Kentucky University, Shandong University

This research presents Neural Dynamic Time Warping (Neural-DTW), a novel framework that learns optimal time warping functions for financial bubble analysis. By replacing traditional dynamic programming with a monotonic attention mechanism, our approach learns to align time series while incorporating multi-dimensional market features. In our exploratory analysis comparing dot-com and AI bubbles, Neural-DTW shows promising potential for reducing alignment error and producing more economically plausible temporal mappings compared to classical methods. This work formalizes time warping as a

learnable optimization problem, bridging neural networks with time series analysis for improved financial pattern recognition.

A Max 0-1 Weighted Log-Rank Test for Time-to-Event Data with Random Delayed Treatment Effect

•*Jiarui Liu, Chenghao Chu, Bingming Yi, Ming-Hui Chen*

Great Bay University

In a cancer clinical trial with an immunotherapy, the treatment often becomes effective after a certain delay time. It is well known that traditional log-rank test no longer provides optimal power in this case. There has been a great interest to search for a more powerful test to detect a truly effective immunotherapy treatment.

In this paper, we propose a new testing method based on the maximum value of a set of 0-1 weighted log-rank test statistics and establish useful theoretical results of the relevant test statistics.

Extensive simulation studies are conducted to examine the empirical performance of the proposed method. An analysis of the real data from a non-small-cell lung cancer clinical trial is also carried out to further demonstrate the usefulness of the proposed test method.

S146: Structured Data Analysis and Dependence Modeling

Flow-Driven Regression for Spatially Heterogeneous Coefficients

Wendong Li

East China Normal University

We propose flow-driven regression (FDR) for estimating spatially heterogeneous regression coefficients with piecewise homogeneous structure and continuous spatial boundaries. The method is built on a new flow-driven penalty (FDP) defined on an oriented planar triangulation, which penalizes local flow imbalance of coefficient differences and is designed to preserve coherent two-dimensional boundary propagation. In contrast to edgewise fusion and tree-based regularization, the proposed construction better reflects the geometry of continuous spatial boundaries while maintaining a compact penalty structure. We show that the associated penalty matrix is nearly full rank on regular lattices, leading to a constrained-lasso formulation with only a small number of linear equality constraints. We further establish nonasymptotic error bounds for estimation and prediction. Numerical studies show that FDR improves boundary recovery and coefficient estimation over existing methods for spatially heterogeneous regression.

Deep Sparse Masks via Optimal Transport

Zihao Tang, • *Yixuan Qiu*

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

Sparse masks, defined as binary vectors with a fixed number of ones, are useful objects in statistical machine learning, with wide applications in feature selection, structure learning, and neural network pruning. However, learning sparse masks from data is an extremely challenging task, as masks are intrinsically discrete objects that are hard to be integrated into modern gradient-based optimization frameworks. In this work, we propose a novel framework that leverages the quadratically regularized optimal transport (QROT) to construct and learn sparse masks. The key idea is to represent a sparse mask by the solution to a QROT problem, which can be parameterized by a continuous parameter vector. To overcome the non-differentiability of the QROT solution, we introduce the Bouligand derivative (B-derivative) in non-smooth analysis as a generalized concept of gradient, and derive an analytic expression for the B-derivative of QROT-based sparse masks. We provide rigorous theoretical guarantees on the validity of the B-derivative, as well as the approximation and statistical errors of the sparse masks learned from the data. Various numerical results demonstrate the effectiveness of QROT-based sparse masks in learning structured models, including variable selection in linear models, weight sparsification in neural networks, and detecting important regions of images in convolutional neural network classification.

Dependence-Adaptive Modelling for Functional Time Series

•*Zerui Guo, Jianbin Tan, Hui Huang*

Sun Yat-sen University, Duke University, Renmin University of China

Functional time series (FTS) have become increasingly available in real-world applications. Research on such data typically focuses on two objectives: curve reconstruction and forecasting, both of which require efficient and accurate dimension reduction. While functional principal component analysis (FPCA) serves as a standard tool, existing methods often commit to rigid representations that may be misaligned with the data-specific dependence structure. Such misalignment can result in either oversimplified or redundant models, leading to suboptimal dimension reduction and degraded predictive accuracy in practice. To address this limitation, we quantify the intrinsic dependence of FTS via a least-squares-type decomposition, projecting the auto-covariance sequence onto a weakly separable reference class. This quantification motivates a set of versatile basis functions, termed optimal functional filters, which we embed into standard FPCA routines to yield dependence-adaptive modeling for FTS with diverse structures. We develop an associated estimation procedure

that operates within a Bayesian framework, incorporates the Whittle likelihood, and remains valid under sparse and irregular sampling. Theoretical properties are established for the proposed methodology, and comprehensive simulations validate its superiority over competing approaches. We further illustrate our method through an application to reconstructing and forecasting daily air pollutant concentration trajectories.

Deconfounding via Profiled Transfer Learning

•*Ziyuan Chen, Yifan Jiang, Jingyuan Liu, Fang Yao*

Peking University, The Pennsylvania State University, Xiamen University, Peking University

Unmeasured confounders are a major source of bias in regression-based effect estimation and causal inference. In this paper, we advocate a new profiled transfer learning framework, ProTrans, to address confounding effects in the target dataset, when additional source datasets that possess similar confounding structures are available. We introduce the concept of profiled residuals to characterize the shared confounding patterns between source and target datasets. By incorporating these profiled residuals into the target debiasing step, we effectively mitigate the latent confounding effects. We also propose a source selection strategy to enhance robustness of ProTrans against noninformative sources. As a byproduct, ProTrans can also be utilized to estimate treatment effects when potential confounders exist, without the use of auxiliary features such as instrumental or proxy variables, which are often challenging to select in practice. Theoretically, we prove that the resulting estimated model shift from sources to target is confounding-free without any assumptions imposed on the true confounding structure, and that the target parameter estimation achieves the minimax optimal rate under mild conditions. Simulated and real-world experiments validate the effectiveness of ProTrans and support the theoretical findings.

S150: The Frontier Research of Statistical Process Quality Control under AI era

A Change Point Control Chart Based on an EWMA-Type Smoothing of the Cross Matching A1 Statistics

Jiujun Zhang

Liaoning University

In the context of smart manufacturing and data-driven quality management, monitoringhigh-dimensional and complex data streams has become a critical requirement for ensur-ing process stability and product reliability. Traditional parametric control charts oftenrely on strict distributional assumptions, which are frequently violated in practical high-dimensional scenarios, leading to compromised monitoring accuracy. Meanwhile, existing nonparametric methods either lack robustness when process parameters are unknown orstnugsle to balance detection sensitivity for smal-to- moderate shits and computational EWMA control chart based on the cross-matching A1 statistic. The proposed chart integrates the advantages of the cross-matching test with the EWMA smoothing technigue which enhances detection nower for suhtle and nersistent nrocess changes It requires no prior knowledge of the underlving process distribution and avoids complex multivariate distribution fitting making it easy to implement in practice. Research has shown that when one or more of the three parameters of potential process distribution, namely position, scale, and skewness, exhibit various types of offsets, the proposed sofectivenessthe proposed method demonstrates excellent performance in external control detection.

A Robust Online Control Chart Based on the Adaptive Conformal Method

•*Miaomiao Yu, Yu Cao*

East China Normal University

In long-distance refined-product pipelines, precise tracking and monitoring of the oil mixing play a central role in guaranteeing oil quality and minimizing economic losses. The monitoring of continuous changes in dynamic data streams are critical for achieving real-time and accurate supervision in industrial processes and for ensuring product quality. Statistical Process Control (SPC), as a well-established online monitoring tool, automatically identifies state changes in the oil-mixing process by constructing control charts. However, existing statistical quality control methods typically determine control limits using Markov or Monte Carlo method, which suffer from low computational efficiency and poor adaptability to high-frequency continuous data streams. Moreover, these methods are inadequate for handling non-normal and non-stationary data. To address these limitations, this paper introduces conformal inference from machine learning into the SPC framework, thereby overcoming the slow convergence and limited applicability of traditional bisection methods. This paper makes two main contributions: First, a quantile-based conformal inference approach is employed to efficiently and rapidly determine control limits. Second, a dual-threshold mechanism is constructed to accommodate real-time minor fluctuations in in-control data caused by external conditions, thereby improving monitoring accuracy. As a result, the proposed framework significantly improves monitoring robustness and accuracy in complex industrial environments. Experimental results based on both simulated data and real-world datasets demonstrate the superiority of the proposed method.

Online Monitoring in Telecommunication Signals via Wasserstein Barycenter Modeling

Wei Zhao

Shanghai Polytechnic University

Ensuring the reliability of telecommunication systems requires effective monitoring of time-evolving signal distributions under dynamic operating conditions. This paper develops a Wasserstein-based modeling and monitoring framework for capturing and detecting distributional changes in signal behavior, such as the Reference Signal Received Power (RSRP). A reference model is first constructed using Wasserstein barycenter regression to describe the temporal evolution of empirical signal distributions. Entropy regularization is incorporated to achieve smooth and stable barycenter estimation, enabling efficient computation through the Sinkhorn algorithm. Based on the reference model, an online monitoring scheme is proposed, where the Wasserstein distance between observed and reference distributions serves as the detection statistic. A smoothed monitoring score is maintained within a moving window, and anomalies are identified when the score exceeds a calibrated control limit determined from historical in-control data. The effectiveness and sensitivity of the proposed framework are validated through comprehensive simulation experiments and a real telecommunication network case study, demonstrating its capability for timely and accurate anomaly detection. The proposed method provides a robust and interpretable approach for detecting subtle and gradual degradation patterns in telecommunication network performance, thereby enhancing system reliability assessment and early fault diagnosis capabilities.

A Bayesian Functional Tensor Framework for Online Monitoring of Continuous-Indexed Dynamic Processes with Application to Meteorological Surveillance

Yifan Li

Nanjing Audit University

Monitoring dynamic processes of multi-way continuous-indexed functional data is increasingly critical in fields such as environmental science and engineering. However, existing online monitoring methods often fall short, as they either assume discrete indices or require vectorizing the data, which discards the intrinsic, high-order spatial structure. In this article, we propose a novel two-stage monitoring framework, the Functional Dynamic Tensor EWMA chart, specifically designed to handle such complex data. In Phase I, we use Functional Bayesian Tucker Decomposition with state-space Gaussian Processes to model historical IC data, introducing an additional mode to represent different processes and imposing a stable prior to quantify regular long-term variations and weak correlations across multiple processes. In Phase II, sequential decorrelation is applied to eliminate temporal correlations, and a multivariate EWMA scheme monitors standardized residuals to detect process shifts. Empirical results from simulations and a real-world meteorological monitoring application show that the proposed chart achieves lower false alarm rates, faster anomaly detection, and stronger robustness than existing methods. The proposed framework is well-suited for industrial and environmental applications requiring real-time monitoring of high-dimensional spatiotemporal dynamic processes.

S144: Statistics and Machine Learning at the Frontiers of Genomics and Genetics

Decoding and AI-Driven Design of Cell Type Specific Cis-Regulatory Elements

Yuan He

Beijing Zhongguancun Academy

Cell type specific gene expression regulation is critical for cell specific biological functions. However, identifying the regulatory grammar that encodes such specificity, and further designing regulatory elements with desired cell type specificity, remains challenging. We constructed a tissue-specific eQTL (expression quantitative trait loci) atlas based on GTEx and identified context specific caQTLs (chromatin accessibility QTLs) by integrating public ATAC-seq datasets, revealing the role of transcription factors in cell type specific regulatory programs. Furthermore, we developed an AI framework GO-CRE (Guided Optimization of Cis-Regulatory Elements). It builds genomic LLM (large language model) with reinforcement learning iterations to directionally optimize and generate cell type specific cis-regulatory elements. MPRA experiments validated the cell type specific activity of the generated CREs in K562 and HepG2 cells, and further revealed how the model learns transcription factor recognition sequences and regulatory grammar during the AI-guided design process. This strategy provides a generalizable framework for designing cell type specific regulatory elements, with potential applications in precision gene regulation.

Spatially Resolved Mapping of Cells Associated with Human Complex Traits

Jian Yang

Westlake University

Depicting spatial distributions of disease-relevant cells is crucial for understanding disease pathology. Here, we present a method, gsMap, that integrates spatial transcriptomics (ST) data with genome-wide association study (GWAS) summary statistics to map cells to human complex traits, including diseases, in a spatially resolved manner. Using embryonic ST datasets covering 25 organs, we benchmarked gsMap through simulation and by

corroborating known trait-associated cells or regions in various organs. Applying gsMap to brain ST data, we revealed that the spatial distribution of glutamatergic neurons (glu-neurons) associated with schizophrenia more closely resembles that for cognitive traits than that for mood traits, such as depression. The schizophrenia-associated glu-neurons were distributed near the dorsal hippocampus, with upregulated calcium signaling and regulation genes, while the depression-associated glu-neurons were distributed near the deep medial prefrontal cortex, with upregulated neuroplasticity genes. Our study provides a method for spatially resolved mapping trait-associated cells and demonstrates the gain of biological insights (e.g., spatial distribution of trait-relevant cells and related signature genes) through these maps.

AI for (Bio) Statistics and (Bio) Statistics for AI

Yiqun Chen

Johns Hopkins University

The rise of AI models has offered much promise to augment expensive human labels with cheap model-generated ones. However, the model predictions could introduce systematic errors, false confidence, and estimates that quietly drift from the truth. This is one half of a broader story about the deep interplay between AI and statistics. In the first half of this talk, we walk through a family of statistical frameworks — drawing on classical ideas from missing data, measurement error, and survey sampling — that use a small number of human-verified labels to correct for the errors in a large pool of noisy AI-generated labels. We show that the core problem reduces to an elegant problem: estimating a conditional expectation, and we connect it to decades of progress in noisy-label and ensemble learning. In the second half, we turn it around and show how statistical principles help AI make better predictions in high-dimensional biological settings. Specifically, we describe how ideas from (bio)statistics improve AI-driven analysis of perturb-seq experiments. Together, these two directions illustrate a productive loop: AI creates new applied opportunities for classical statistical tools, and relevant statistical thinking makes AI work better in the sciences.

Bayesian Rhythmic Model for Jointly Detecting Circadian Biomarkers and Predicting Molecular Circadian Time in Human Post-Mortem Brain Transcriptome

George C. Tseng

University of Pittsburgh

Transcriptomic circadian analysis in human post-mortem brain has provided an unprecedented opportunity to decipher in vivo molecular circadian rhythms in different human brain regions that are known to play essential roles in aging and psychiatric disorders. Detecting circadian biomarkers is often the first objective in the statistical analysis. Confounded with the task is the common issue that the true molecular circadian clock time within a human subject is usually inconsistent with the record, which could come from observation or recording errors, or the intrinsic biological variation. In the literature, many methods have been proposed for detecting rhythmic biomarkers or for predicting true molecular circadian time, respectively. To date, no method is developed to achieve both objectives simultaneously. In this paper, we propose a Bayesian model for simultaneous Circadian marker detection and molecular circadian Time estimation (BayCT). The model is further extended to repeated measure of multiple tissues or brain regions. We adopt Von Moses prior distribution for angular data with slice sampling and reversible jump sampling in the Markov chain Monte Carlo (MCMC) procedure for Bayesian inference. We demonstrate the method by extensive simulations and two applications in 12 tissues in mouse brain and three brain regions in human. The result shows superior performance in both statistical objectives with large margin in both circadian marker detection and circadian time detection, as well as the advantage of integrating multiple brain regions.

S001: Advances in Statistical Methods for Genetic and Genomic Studies

Integrative Analysis and Regulatory Inference for Spatial Multi-Omics Data

Jishuai Miao, Jinzhao Li, Jingxue Xin, Jiajuan Tu, Muyang Ge, Ji Qi, Xiaocheng Zhou, Ying Zhu, Can Yang, •Zhixiang Lin

The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong, Hubei University of Technology, The Chinese University of Hong Kong, The Chinese University of Hong Kong, The Chinese University of Hong Kong, Fudan University, The Hong Kong University of Science and Technology, The Chinese University of Hong Kong

New spatial multi-omics technologies, which jointly profile transcriptome and epigenome/protein markers for the same tissue section, expand the frontiers of spatial techniques. Here, we introduce MultiGATE, which utilizes a two-level graph attention auto-encoder to integrate the multi-modality and spatial information in spatial multi-omics data. The key feature of MultiGATE is that it simultaneously performs embedding of the spatial pixels and infers the cross-modality regulatory relationship, which allows deeper data integration and provides insights on transcriptional regulation. We evaluate the performance of MultiGATE on spatial multi-omics datasets obtained from different tissues and platforms. Through effectively integrating spatial multi-omics data, MultiGATE both enhances the extraction of latent embeddings of the pixels and boosts the inference of transcriptional regulation for cross-modality genomic features.

A Flexible and Unified Framework for Single- and Multi-Outcome Mendelian Randomization Using Summary Statistics

Lin Chen

University of Chicago

Mendelian randomization (MR) is widely used to evaluate the causal effects of complex trait exposures on disease outcomes. Recently, MR has been increasingly applied to molecular traits, such as gene expression, as exposures to map risk genes. However, transcriptome-wide MR (TWMR) faces unique challenges. The number of available cis-QTLs as instrumental variables (IVs) is often limited, and horizontal pleiotropy is pervasive, violating core MR assumptions and compromising inference validity. We introduce FusioMR, a robust MR framework tailored for molecular trait exposures while also applicable to complex trait exposures. Our single-outcome model, FusioMRs, incorporates gene-region-specific empirical priors informed by the number and strength of QTLs, linkage disequilibrium, and effect size consistency. It uses sampling-based inference to improve robustness when instruments are limited. Our multi-outcome model, FusioMRm, is motivated by the observation that many complex diseases have correlated diseases, subtypes, or comorbidities, which could be affected by shared or correlated exposures. FusioMRm jointly analyzes two correlated outcomes, leveraging shared IVs and pleiotropic effects of shared/correlated exposures to improve invalid IV detection and enhance estimation precision and power, particularly for underpowered outcomes. We applied FusioMRs to identify cell-type-specific gene expression traits associated with Alzheimer's disease using single-cell eQTL and GWAS summary data. We applied FusioMRm to detect alternative polyadenylation events affecting atrial fibrillation and ischemic stroke. We also applied FusioMRm to estimate the causal effect of low-density lipoprotein on ischemic stroke in South Asian populations by borrowing information from European ancestry data. These applications highlight the generalizability of FusioMR for both molecular and complex trait exposures.

Deep Learning-Based Hierarchical Community Detection for High-Dimensional Gene Regulatory Networks

Jarred Kvamme, Jordan Davis, •Audrey Fu

University of Idaho, New Jersey Institute of Technology, Wayne State University

Reconstructing genome-wide gene regulatory networks (GRNs) from genomic data is challenging due to high dimensionality and complexity. We propose a hierarchical model with three layers: individual genes at the bottom, gene communities in the middle, and communities of communities at the top, revealing patterns at different scales.

We developed DeepHCD, a deep learning algorithm using a graph attention autoencoder to learn low-dimensional embeddings and infer community structures top-down. DeepHCD minimizes a multitask loss function encompassing graph reconstruction, attribute reconstruction, clustering, and modularity, requiring only rough upper bounds for community numbers at each level.

Simulations across diverse network types demonstrate DeepHCD's superior performance in detecting middle-layer communities using homogeneity and completeness metrics. Applied to single-cell regulon activity data (243 regulons, 30,000+ cells), DeepHCD outperforms existing methods, producing clearer community structures with the highest intra-group correlations.

Building a Causality-Aware Single-Cell RNA-seq Foundation Model via Context-Specific Causal Regulation Modeling

Kaichen Xu, Yuxuan Shen, Maocheng Zhu, Zhenke Duan, Jin Liu, •Xiaobo Sun

Emory University, Georgia Institute of Technology, Zhongnan University of Economics and Laws, Zhongnan University of Economics and Laws, The Chinese University of Hong Kong, Shenzhen, Emory University

Single-cell RNA-seq Foundation models (scFMs) learn transferable gene and cell representations for diverse downstream

tasks, among which inferring cell-specific gene regulatory networks (csGRNs) is particularly valuable for elucidating heterogeneous regulatory mechanisms underlying development and disease. However, existing scFMs largely overlook causal gene relationships and remain compu- tationally infeasible to infer csGRNs at atlas scale, resulting in embeddings with limited relational semantics and restricting the discovery of fine-grained regulatory mechanisms. Here, we introduce scCAFM, a causality-aware scFM that efficiently infers causal csGRNs within a mixture-of-experts- mediated latent causal factor space via a structure foundation module, and subsequently learns causality-aware, contextual gene and cell embeddings informed by the inferred csGRNs via an embed- ding foundation module. Across diverse benchmarks, scCAFM achieves superior accuracy, scalability, and biological interpretability in csGRN inference. Moreover, we demonstrate that scCAFM-derived embeddings are enriched in causal relational semantics and consistently outperform state-of-the-art scFMs across various downstream applications, particularly in csGRN-based tasks such as in silico prediction of perturbation-induced cell fate transitions.

S003: Advanced Methods for Dependence, Signals and Spatio-Temporal Detection

Double Validity of Robust Standard Errors under Complex Dependence

•*Wen Zhou, Lihua Lei, Peng Ding, Zifeng Zhang*

New York University

Regression is a workhorse for analyzing data from randomized experiments, adaptive trials, and related designs, where valid inference typically relies on {\it robust standard errors}. These adjustments are usually justified as correcting dependence in unobserved errors, such as dependence induced by shocks, interference, or hidden common causes, even though such dependence is rarely observable or reliably modeled in practice. By contrast, the dependence structure of the regressor of interest, most notably treatment assignment, is often known, testable, and engineered by design. Motivated by this asymmetry, we develop a general asymptotic framework for regression inference under unknown and potentially unstructured dependence and establish a dual validity principle: various robust t-statistics remain asymptotically valid if the correlation structure encoded by the chosen standard-error adjustment is correct for either the regressor of interest or the unobserved errors, even when the other component is misspecified. Establishing this predictor-error symmetry is mathematically nontrivial and, to our knowledge, has not been formalized in the classical robust-inference literature despite its centrality to empirical practice. For example, cluster-robust inference remains valid when either errors are clustered or assignment is clustered. The result reframes robust standard errors as design-aligned devices: when error dependence is uncertain, robust adjustments should be chosen to match the known assignment correlation, rather than used as a catch-all correction for unspecified error structure.

Invariance Theory for Adaptive Signal Detection

•*Mengjiao Tang, Yao Rong, Augusto Aubry, Antonio De Maio*

Yunnan University, Yunnan University, University of Naples Federico II, University of Naples Federico II

Adaptive signal detection in complex environments faces major challenges from unknown nuisance parameters, structured disturbance, and false-alarm control. This talk introduces invariance theory as a principled framework for addressing these issues. Through suitable group transformations, composite detection problems can be reduced to maximal invariant representations, which enable nuisance-parameter compression, reveal the essential detection-relevant information, and support rigorous constant-false-alarm-rate analysis. This framework also guides the construction of structured adaptive detectors in Gaussian and non-Gaussian settings. The resulting methodology provides an interpretable and robust foundation for detector design in heterogeneous and sample-limited environments.

Likelihood Ratio Tests for Poisson Graph with Directed Acyclicity Constraints

Xin Liu

Shanghai University of Finance and Economics

Count-valued data are pervasive in modern applications, such as sequencing reads, incident tallies, and traffic flows, yet most procedures for learning directed acyclic graphs (DAGs) remain tailored to Gaussian or otherwise continuous measurements. We develop a Poisson DAG framework that models conditional intensities via Poisson regressions and enforces acyclicity through a tractable non-convex constraint. Building on this estimator, we introduce two likelihood–ratio tests: a linkage test for the presence of directed edges and a pathway test for sets of edges. We derive the sampling distributions of the proposed statistics under both the null and alternative hypotheses. When the number of edges tested is fixed, the statistic is asymptotically χ^2; when the dimension grows, a normal approximation applies. We also present an alternating-direction algorithm with closed-form updates that scales to high-dimensional graphs. In simulations across diverse graph topologies and signal regimes, the procedure recovers directions accurately and controls type-I error. Analyses of real count data further demonstrate the ability of the proposed method to uncover interpretable linkages and pathways. Together, these contributions provide a principled and practical toolkit for

causal discovery with count data.

Supervised Spatial Metric Learning for Clustering and Prediction

Xinyue Zhang, ⋆ Hong Gu, Abdrew Irwin, Tobias Kenney

Dalhousie University, Dalhousie University, Dalhousie University, Dalhousie University

Spatial patterns and dependencies play a central role in statistical modeling and inference across many domains. We propose a novel supervised spatial metric learning approach based on Random Forests to quantify similarity between locations by capturing response-driven spatial structure. The method augments spatial coordinates to allow flexible partitioning of the space into irregular subregions. Pairwise similarity is then defined as the proportion of trees in which two locations fall into the same terminal node, conditional on other predictors.

The resulting similarity matrix is used to derive eigen-scores and cluster labels, which are incorporated into a range of predictive models, including geographically weighted regression (GWR), generalized additive models (GAM), Random Forest, XGBoost, spatial and non-spatial varying coefficient models, and the Spatial Durbin Model (SDM). Simulation studies and real data applications demonstrate that the proposed approach effectively captures complex spatial structure, yields meaningful clustering, and substantially improves predictive performance across models.

S015: Advances in Predictive and Integrative Modeling for Complex Health Data

Building Trust in Clinical AI: An Explainable Decision Support System for Chronic Kidney Disease

Krishna Mridha, Ming Wang, ⋆ Lijun Zhang

Case Western Reserve University, Case Western Reserve University, Case Western Reserve University

Chronic Kidney Disease (CKD) is a significant global public health issue, affecting over 10% of the population. Timely diagnosis is crucial for effective management. Leveraging machine learning within healthcare offers promising advancements in predictive diagnostics. We developed a Web-Based Clinical Decision Support System (CDSS) for CKD, incorporating advanced Explainable AI (XAI) methods, specifically SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). The model employs and evaluates multiple classifiers: KNN, Random Forest, AdaBoost, XGBoost, CatBoost, and Extra Trees, to predict CKD. The effectiveness of the models is assessed by measuring their accuracy, analyzing confusion matrix statistics, and the AUC. AdaBoost achieved a near perfect accuracy rate. Except for KNN, all classifiers consistently reached perfect precision and sensitivity. Additionally, we present a real-time web-based application to operationalize the model, enhancing trust and accessibility for healthcare practitioners and stakeholder.

Variational Bayesian Inference for Poissonian Image Restoration

Jinling Liu, ⋆ Chao Wang, Youwei Wen

Southern University of Science and Technology, Southern University of Science and Technology, Hunan Normal University

In this work, we propose a variational Bayesian framework for Poisson image denoising that integrates a nonlinear data-fidelity term with total variation and nonlocal low-rank regularization. Using quadratic approximations, the complex posterior induced by a non-Gaussian likelihood and non-smooth priors is embedded into a variational Bayesian framework, enabling joint inference of the image and regularization parameters. This framework adaptively balances total variation and nonlocal low-rank priors, while providing uncertainty quantification for the estimates. The convergence of the proposed model is theoretically guaranteed, and the high-probability error bound is established. Numerical experiments demonstrate the effectiveness and robustness of the proposed method. Furthermore, the proposed framework naturally extends to image denoising under Gaussian, Poisson–Gaussian, and Cauchy noise distributions.

BioBMS: A Knowledge-Guided Dynamic Bayesian Screening and Dual-Prior Validation Framework for Robust Biomarker Discovery

He Mu

Xi'an Jiaotong-Liverpool University

Background:
High-dimensional metabolomic biomarker discovery often suffers from severe collinearity and historical publication bias, leading to prognostic failures in diverse real-world populations. Although abundant historical literature exists, systematically integrating it into prospective screening without overwhelming empirical clinical evidence remains challenging.

Methods:
We propose BioBMS, a Knowledge-Guided Dynamic Bayesian Screening framework, applied to a UK Biobank prospective cohort across seven neurological disorders. BioBMS translates historical literature evidence (via BioBERT semantic similarity and impact factors) into Bayesian informative priors. To handle collinear noise, we instituted a knowledge-guided smart pruning strategy. Furthermore, we developed a Dual-Prior Adversarial Bayesian Validation (DPABV) mechanism, utilizing Bayesian Spike-and-Slab regression to contrast an informed model against

an uninformative null baseline.

Results:

While literature-informed priors acted as a safe regularizer, they yielded modest predictive gains over the data-driven baseline (e.g., AUC improved from 0.824 to 0.825 for Alzheimer's disease). Crucially, the DPABV framework successfully quantified latent

"prior-data conflicts". We observed robust prior-empirical concordance in Alzheimer's disease (R=0.706, P=0.023) and Vascular Dementia, but pervasive decoupling in conditions like Epilepsy (R=-0.031, P=0.933), highlighting systemic mismatches between heavily-published pathways and prospective utility. Following rigorous multi-seed stochastic filtering, BioBMS isolated extremely sparse, pathologically grounded signatures from high-dimensional noise, notably Glucose for Vascular Dementia (PIP>0.95) and Acetone for Epilepsy (PIP=0.765).

Conclusion:

BioBMS effectively isolates robust metabolic biomarkers by balancing historical knowledge with empirical likelihood constraints. Beyond predictive optimization, it serves as a critical diagnostic pipeline to expose literature-data decoupling, preventing the translation of publication bias into biomarker discovery.

S016: Advances in Statistical Inference for Complex and High-Dimensional Models

Integral Probability Metric-Guided CUSUM-Net for Nonparametric Changepoint Detection

Zhou Yu

East China Normal University

We propose CUSUM-Net, a nonparametric method for changepoint detection based on integral probability metrics (IPMs) and deep neural networks. Our approach learns a critic function by maximizing an aggregate CUSUM objective over candidate changepoints, thereby linking changepoint detection to two-sample IPM optimization. The learned critic induces a one-dimensional representation on which changepoints are localized by a classical CUSUM scan. Unlike parametric procedures, CUSUM-Net accommodates complex, high-dimensional distributional changes and applies to a range of data modalities, including Euclidean data, symmetric positive-definite matrices, images, and graphs. We establish excess-risk bounds for the learned critic under Hölder smoothness assumptions, with faster rates when the data exhibit low-dimensional manifold structure, and we derive corresponding changepoint localization guarantees. Numerical experiments demonstrate the flexibility and effectiveness of the proposed method.

Neural Wasserstein Two-Sample Tests

◆ *Xiaoyu Hu, Zhenhua Lin*

Xi'an Jiaotong University, National University of Singapore

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 1-Wasserstein distance, which we call the neural Wasserstein test. 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.

Structural Change Detection in Dynamic Systems

◆ *Wei Zhang, Fang Yao*

Fudan University, Peking University

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.

Semiparametric M-estimation with Overparameterized Neural Networks

Shunxing Yan

Peking University

Semiparametric regression has played a central role in statistics, integrating the expressive power of deep neural networks with interpretable statistical inference on parameters of interest. Despite the success of classical semiparametric method/theory, the root-n-asymptotic normality and inference of the finite-dimensional parameter estimator in this context remain challenging, mainly due to nonlinearity and potential tangent space degeneration. In this work, we introduce a foundational framework for semiparametric M-estimation using overparameterized neural networks. Our approach guarantees desirable tangent space behavior and we analyze the statistical properties of algorithmic estimators with optimization guarantee. Nonparametric convergence and parametric asymptotic normality for general loss are established, enabling valid statistical inference. Furthermore, from high-level perspectives, this work introduces a approach for analyzing nonparametric tangent behavior in statistical inference works involving DNNs and also provides a new viewpoint for deep learning on the importance of network tangent behavior through the lens of semiparametric statistics.

S026: AI-Driven Genomics and Precision Medicine: From Prediction to Personalization

MethylFM: A DNA Methylation Foundation Model for Modeling Epigenomic Regulatory Dynamics

Limeng Pu, ◆ *Xiang Chen*

St. Jude Children's Research Hospital, St. Jude Children's Research Hospital

DNA methylation regulates gene expression, differentiation, and disease, making it a key target for computational modeling in precision medicine. To integrate high-resolution whole-genome bisulfite sequencing (WGBS) data into a unified analytical framework, we developed MethylFM, a transformer-based foundation model that captures context-aware methylation patterns and supports multiple downstream tasks.We trained on the BLUEprint Epigenome dataset, comprising single-base WGBS profiles across diverse blood cell types with matched histone modification and transcriptome data. By focusing on ±100 CpG sites around transcription start sites (TSS), MethylFM targets regulatory regions central to gene expression and chromatin dynamics. Built on a BERT-style transformer with rotary positional embeddings and a masked-value prediction objective, it learns robust representations capturing both local and long-range dependencies in methylation.MethylFM demonstrated versatility across three downstream applications. For CpG-level imputation, it reconstructed high-resolution WGBS profiles from 450k array data with strong accuracy ($R^2 > 0.6$, MAE < 0.15). When benchmarked against METHimpute, evaluation was restricted to two samples due to METHimpute's intensive runtime; nonetheless, MethylFM achieved slightly higher accuracy ($R^2 = 0.518$ vs. 0.513), underscoring both precision and computational efficiency. In TSS-level H3K27ac prediction, the model reached $R^2 = 0.614$, matching the state-of-the-art M2A ($R^2 = 0.617$) and highlighting its capacity to infer promoter activity directly from DNA methylation. Finally, sample-level clustering based on predicted H3K27ac profiles accurately recapitulated hematopoietic lineages, surpassing experimental H3K27ac data (Silhouette = 0.47 vs. 0.30) and approaching RNA-seq-derived clustering performance (Silhouette = 0.51), demonstrating that MethylFM captures biologically meaningful epigenetic structure. Together, these results establish MethylFM as a generalizable and efficient framework for epigenomic modeling, enabling cost-effective methylation imputation, promoter activity prediction, and cellular identity characterization to advance biomarker discovery and precision medicine.

Toward Robust Genomic Prediction Across Asian Ethnicities: Insights from Machine Learning Approaches

◆ *Qiao Fan, Tingtong Liu, Xiaomin Liu, minhao Yao*

Duke-NUS Medical School, Duke-NUS Medical School, Duke-NUS Medical School, Duke-NUS Medical School

Genetic prediction for common diseases has been extensively explored; however, some issues remain in optimizing prediction models across multi-ethnic populations.

Here, we investigated genetic prediction for age-related macular degeneration using multi-ethnic Asian cohorts comprising 8621 cases and controls with 6 million genetic variants. Machine learning models, including LASSO regression, random forest, XGBoost, and neural networks, were applied to evaluate predictive performance. We examined the impact of linkage disequilibrium (LD) structures by varying r^2 thresholds for clumping and pruning, assessed variant selection strategies, and explored cohort-specific fine-tuning within individual ethnic groups. Model performance was compared with polygenic risk score (PRS) approaches. Our findings highlight key considerations and practical strategies for implementing genetic prediction in multi-ethnic datasets with small to moderate sample sizes.

AdaBoost-rDNN: A Boosting Ensemble of Deep Neural Networks with Regularization for Disease Outcomes

◆ *Yaqing Xu, Mengni Zhang, Kuangnan Fang, Shuangge Ma*

Shanghai Jiao Tong University School of Medicine, Xiamen University, Xiamen University, Yale School of Public Health

Predicting disease outcomes is a complex task due to the interplay of genetic, environmental, and clinical factors. Existing research has demonstrated the effectiveness of ensemble learning in biomedical and related fields. This study introduces AdaBoost-rDNN, an ensemble deep learning approach that incorporates gene-environment (G-E) interactions and uses regularization to enhance prediction accuracy for disease outcomes. The proposed method leverages boosting to combine weak learners based on deep neural networks, which are specifically designed to handle high-dimensional interactions with hierarchical structure. By applying regularization, the base learner filters out irrelevant information, ensuring that only significant genetic and G-E interactions remain. Simulation studies demonstrate the effectiveness of AdaBoost-rDNN. The proposed approach consistently outperforms traditional machine learning techniques, showcasing its potential in improving cancer prediction. Overall, this study offers a novel ensemble deep learning approach that incorporates G-E interactions, providing improved prediction for disease outcomes.

S024: Advancing Complex Dependent Data Analysis

Mixture of Dynamic Experts: Traffic Flow

◆ *Zhengwei Liu, Sheng Jiang*

Nanjing Normal University, The Chinese University of Hong Kong, Shenzhen

Urban traffic flow data exhibit heterogeneity, temporal dynamics, and spatial variation. We propose a Mixture of Dynamic Experts model for count-valued traffic data in Beijing, where observations follow Poisson distributions with latent mixture components governed by a multinomial allocation. The log-intensity of each component is modeled via a Dynamic Linear Model (DLM) to capture time evolution, while geographic covariates are incorporated to reflect spatial heterogeneity. The proposed framework provides a flexible approach for modeling complex traffic patterns and enables improved inference and understanding of urban traffic systems.

Robust Estimation and Bootstrap Testing for Self-Excited Threshold Poisson Autoregressive Model

Xinyang Wang

Liaoning University

Self-exciting threshold Poisson autoregressive (SETPAR) model is widely used for modeling piecewise-linear dynamics in integer-valued time series. However, the classical maximum likelihood estimator (MLE) is notoriously sensitive to outliers, often leading to severely biased parameter estimates and structural misidentification. To address this deficiency, we propose a robust Mallows quasi-likelihood estimation (MQLE) procedure. The asymptotic of the proposed estimator is establised. Moreover, we develop a supremum-type robust score test and employ a parametric bootstrap procedure to detect the presence of threshold effects. Monte Carlo simulations demonstrate the robustness of our method, and a real data example further highlights its empirical effectiveness.

Several Types of Binomial Autoregressive Models with Endpoint Inflations and Their Statistical Inference

◆ *Jie Zhang, Xiaojing Fan, Kai Yang, Yao Kang*

Changchun University of Technology, Changchun University of Technology, Changchun University of Technology, Xi'an Jiaotong University

Bounded count time series arise extensively in practical scenarios such as toxicology experiments, quality control, and voting behavior, with binomial autoregressive models and their extensions serving as standard analytical tools. However, real-world data often exhibit features deviating from standard binomial distributions, including overdispersion and endpoint inflation. Although zero inflation corresponding to the left endpoint has been thoroughly studied, systematic modeling and testing of right-endpoint inflation in bounded count time series have long lacked attention, and there is a notable absence of effective tools capable of simultaneously characterizing endpoint inflation and individual dependence structures. Consequently, we formally incorporate right-endpoint inflation into the modeling framework of the first-order binomial autoregressive (BAR(1)) process for bounded count time series. We propose three extended BAR(1) models with endpoint inflation, and systematically investigate the stochastic properties of the new models. The estimation of parameters is studied by using conditional maximum likelihood and quasi-maximum likelihood methods, and a binomial right-endpoint inflation index is constructed for testing whether a given dataset exhibits endpoint-inflated characteristics relative to a BAR(1) process. Two empirical applications demonstrate that the proposed models offer significant advantages in fitting bounded count time series with endpoint inflation. Furthermore, based on a generalized binomial thinning operator that accommodates dependence among individuals, we introduce a novel endpoint-inflated binomial autoregressive process that effectively characterizes count time series data with both endpoint inflation and individual dependence structures. A thorough analysis of the stochastic properties of the model is provided, along with conditional maximum likelihood and conditional least squares estimators and their asymptotic properties. A joint test

for endpoint inflation and overdispersion against the BAR(1) process is constructed. Numerical simulations and a comparative case study against existing benchmark models confirm the effectiveness of the proposed approach.

A Composite Framework for Correcting Under-Reporting in Count Data with Dependence

Yixuan Fan

Tianjin University

Count data are often subject to under-reporting, especially in infectious disease surveillance. However, this issue has received much less attention when imperfect reporting coincides with serial dependence and overdispersion. In this paper, we propose a framework for under-reported count data with INGARCH-type dependence that separates latent incidence dynamics from the reporting mechanism with time-varying reporting probabilities. Relying only on an informative prior distribution for the mean reporting rate, we develop a Bayesian inference scheme that jointly estimates latent counts, autoregressive parameters, reporting probabilities, and the overdispersion parameter via efficient data augmentation and marginalization techniques. Numerical simulations demonstrate improved parameter recovery, calibrated uncertainty quantification, and superior predictive performance compared with models that ignore under-reporting.

S034: Conditional Modeling and Causality: Generative Models for Learning and Integration

Semi-Supervised Conditional Diffusion via Label Augmentation

◆ *Jin Su, Yuan Gao, Yong Zhou, Jian Huang*

The Hong Kong Polytechnic University, Nankai University, East China Normal University, The Hong Kong Polytechnic University

Conditional diffusion models have become a powerful and flexible framework for learning complex conditional distributions from labeled data. In practice, however, acquiring high-quality labels is costly and time-consuming, leaving large volumes of unlabeled data unused. To address this, we introduce labelaugmented conditional diffusion (LACD), a simple and effective approach that incorporates unlabeled examples by assigning them a designated trivial label and performing joint denoising score matching over the augmented dataset. We provide sufficient conditions guaranteeing populationlevel identifiability of the target conditional distribution under this scheme. Moreover, we establish rigorous statistical guarantees: the sampling distribution produced by LACD converges strictly faster than the purely supervised estimator in total variation distance, and at least as fast in Wasserstein1 distance. Extensive experiments on synthetic, image, and tabular benchmarks corroborate our theory and show substantial gains in sample efficiency and generative performance compared with the purely supervised estimator.

Causal Effect Heterogeneity Estimation Using Summary Statistics

Xingjie Shi

East China Normal University

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.

From Conditional to Unconditional Independence: Testing Conditional Independence via Transport Maps

◆ *Chenxuan He, Yuan Gao, Liping Zhu, Jian Huang*

Renmin University of China, Nankai University, Renmin University, The Hong Kong Polytechnic University

Testing conditional independence between two random vectors given a third is a fundamental and challenging problem in statistics, particularly in multivariate nonparametric settings due to the complexity of conditional structures. We propose a framework for testing conditional independence using transport maps. At the population level, we show that two well-defined transport maps can transform the conditional independence test into an unconditional independence test, thereby substantially simplifying the problem. These transport maps are estimated from data using conditional continuous normalizing flow models. Within this framework, we derive a test statistic and prove its asymptotic validity under both the null and alternative hypotheses. A permutation-based procedure is employed to evaluate the significance of the test. We validate the proposed

method through extensive simulations and real-data analysis. Our numerical studies demonstrate the practical effectiveness of the proposed method for conditional independence testing.

Conditional Augmentation Enables Effective Use of Synthetic Data in Diffusion Models

◆ *Xueyu Zhou, Yuan Gao, Jian Huang*

The Hong Kong Polytechnic University, Nankai University, The Hong Kong Polytechnic University

Synthetic data generated by foundation models has recently emerged as a promising resource for acquiring pre-trained knowledge and improving data efficiency, especially in scenarios where real data is limited. However, directly incorporating synthetic data often introduces distributional bias and can even result in model collapse. To address these challenges, we propose Conditional Augmentation with Synthetic Data (CASD), a framework guided by three core principles: (a) No harm and no bias: The use of synthetic data should neither degrade model performance nor introduce distributional bias; (b) Positive utility: Synthetic data should enhance model performance; (c) Broad adaptability: The approach should be applicable to synthetic data from diverse sources without requiring case-specific modifications. CASD leverages both real and synthetic data by conditioning on their source labels, treating them as related but distinct domains. This design enables the model to harness large-scale synthetic data to strengthen representation learning, while mitigating bias by focusing on the conditional distribution. During sampling, the model utilizes the enhanced representation function to extract useful information from synthetic data, but fixes the source label to the real domain, ensuring consistency with the target distribution. Experimental results demonstrate that CASD can effectively utilize synthetic data from various foundation models, consistently improving both the quality and diversity of generated images without inheriting distributional bias.

S047: Frontiers in Discrete-Valued Time Series Modeling

Z-Valued Smooth Transition GARCH Models: Specification and Testing

Fukang Zhu

Jilin University

We introduce a new class of nonlinear models known as the Z-valued smooth transition GARCH model, designed to accommodate Z-valued time series that display asymmetric, nonlinear and highly persistent volatility. The article outlines the maximum likelihood estimation procedure and establishes its consistency and asymptotic normality of the estimated parameters. Three types of tests are studied, including sup-type linearity test, score-based goodness-of-fit test, and residual-based mixed portmanteau diagnostic checking test. The asymptotic properties of these three test statistics are established. To address the computationally complex problems of estimation, the parameterization of the smooth transition function and the optimization algorithm for the estimation procedure in numerical simulations are discussed. The effectiveness of the tests is demonstrated through numerical simulations, and crime and exchange rate datasets are analyzed to showcase the superior performance of the proposed model.

Inference for Z-Valued INGARCH Models Allowing for Negative Autocorrelation

Yuqing Zhang, Baichuan Lin, ◆ *Yao Kang*

Xi'an Jiaotong University, Xi'an Jiaotong University, Xi'an Jiaotong University

Z-valued versions of integervalued generalized autoregressive conditional heteroscedastic (INGARCH) models are widely used to analyze Z-valued time series, which consist of discrete, quantitative observations from $Z=\{...,-1,0,1,...\}$ and frequently arise in economics and finance. However, the attainable range of autocorrelation function (ACF) values for the existing Z-valued INGARCH models is restricted, and negative ACF values are usually not possible. To better fit Z-valued time series, this talk introduces two Z-valued INGARCH models based on the reparameterized Skellam distribution and softplus link function. The statistical inference for the proposed models is investigated. Applications to financial datasets with positive and negative autocorrelation demonstrate the effectiveness of the proposed models.

Quantile Regression and Inference for Count Network Time Series

Danshu Sheng

Harbin Institute of Technology

Quantile regression is fundamental in modern statistics, yet a tractable framework for conditional quantile modeling and inference for dependent count processes remains largely unavailable, especially for count network time series.

The obstruction is structural: existing count time series formulations typically do not yield a direct and tractable representation of conditional quantiles, while discreteness further complicates quantile-based inference through mass points and the nondifferentiability of the check loss.

We resolve this gap by introducing a continuous latent layer for count processes, under which latent conditional quantiles become

directly modelable under temporal and network dependence.

This yields a general framework for conditional quantile modeling, estimation, and inference in a broad class of count network time series models.

A ceiling-based mapping then transfers the latent conditional quantiles back to the conditional quantiles of the original count process.

Under high-level regularity conditions, we establish a unified theory for quantile-based estimation, inference, and diagnostic checking for both fixed and growing network size, with classical count time series nested as the special case where the network size is one.

We further illustrate the theory through a fully specified representative model, namely the count network quantile vector autoregression, together with simulations and an application to California earthquake magnitude class data.

Forecasting Natural Disaster Frequencies Using Nonstationary Count Time Series Models

◆ *Jian Pei, Yang Lu*

Beijing University of Civil Engineering and Architecture, Concordia University

Because of the climate change, the frequency of natural disaster might be evolving. This could mean an increasing expected risk, and/or more and more uncertainties. In this paper, we identify three potentially suitable models, that are the nonstationary INGARCH(1, 1), the nonstationary INAR(p), and the state-space model of Harvey and Fernandes. We derive properties of their long-run behavior, discuss their link and differences, and assess their suitability for Canadian climate event data. We show that first, the Harvey–Fernandes model or INGARCH(1, 1) model often provides better fit and better prediction performance. Second, the long-run prediction of these models can differ substantially, highlighting model uncertainty.

S048: Frontiers in Learning and Optimization: From GenAI to Dynamic Operations

Robust Sequential Experimental Design for A/B Testing

Qianglin Wen, Xiangkun Wu, Chengchun Shi, Ting Li, Niansheng Tang, ◆ *Yingying Zhang, Hongtu Zhu*

Yunnan University, Zhejiang University, East China Normal University

Experimental design has emerged as a powerfulapproach for improving the sample efficiency ofA/B testing, yet existing designs rely criticallyon correctly specified models. We study robustsequential experimental design under model misspecification and develop a unified frameworkthat covers both contextual bandit and dynamicsettings. Theoretically, we prove that our design bounds the worst-case mean squared errorof the estimated treatment effect. Empirically,we demonstrate the effectiveness of the proposedapproach using synthetic and real-world datasetsfrom a leading technology company.

Statistical Inference in an Interactive Learning Paradigm

Yuchen Wu, ◆ *Kangjie Zhou, Weijie Su*

Cornell University, Columbia University, University of Pennsylvania

The proliferation of generative artificial intelligence has given rise to an \emph{interactive learning} paradigm, where model parameters are continuously updated using not only data generated by natural processes, but also synthetic outputs produced by other models.

This paradigm introduces two major challenges: (1) training data are no longer drawn exclusively from the target population, undermining a core assumption of classical statistical learning theory, and (2) model training processes become inherently correlated, as models influence one another through repeated exposure to each other's synthetic outputs.

Establishing reliable statistical inference in such interactive environments therefore remains an important open problem.

In particular, there is growing concern about \emph{model collapse}, a phenomenon in which the performance of generative models progressively degrades as they are trained on synthetic data produced by earlier model generations.

In this talk, we study the behavior of generative models under general interaction patterns.

We formalize these interactions using directed graphs and show that the occurrence of model collapse depends critically on the graph's topology.

Within this framework, we derive an explicit necessary and sufficient condition characterizing when model collapse occurs.

Our analysis covers both finite-sample results for linear regression and asymptotic guarantees for general M-estimators.

We further validate our theoretical findings through extensive numerical experiments.

Integrating High-Dimensional Censored Data Under Privacy Constraints Via Localized Computations

◆ *Xin Ye, Yanyan Liu, Bingyao Huang*

Guangdong University of Finance and Economics,Wuhan

University, Guangdong University of Technology

Limited sample size and censoring inherently limit the statistical efficiency of high dimensional data analysis. While integrating data from multiple sources can enhance estimation efficiency, concerns remain regarding data privacy breaches and between-site heterogeneity. In this paper, we propose a privacy-preserving approach to integrate the high-dimensional right-censored data with source-level heterogeneity. The proposed method is based on the local computation strategy: each site can obtain an integrative estimation based on its local full dataset and the summary statistics from other sites. For each party, this strategy not only meets the data privacy constraints but also maximizes its local data's utilization. Moreover, we introduce a refined procedure for practical use to avoid the shrinkage of the local covariate effect that is unique across all sites. Theoretical results of the proposed estimates including consistency, asymptotic normality and efficiency gains are attained. Simulation experiments demonstrate its superiority over the integrative methods relying solely on summary statistics and the local estimations. The application to multi-source clinical data of ovarian cancer further verifies its practical effectiveness.

S049: Frontiers in Methods for Complex-structured Survival Data

Distributed Privacy-Preserving Group Inference for High-Dimensional Generalized Linear Models

Yingrui Wang, Xuerong Chen, Liuquan Sun, Jian Huang, Min-Qian Liu, ◆ *Dongxiao Han*

Nankai University, Southwestern University of Finance and Economics, Chinese Academy of Sciences and School of Mathematical Sciences, The Hong Kong Polytechnic University, Nankai University, Nankai University

This article introduces a novel method for distributed, differentially private group inference in the high-dimensional generalized linear model. We consider a setting with an untrusted server and a group of trusted, data-holding clients. Each client first constructs a local non private test statistic based on a weighted quadratic function of the regression sub-vector corresponding to the group, and the debiasing and re-weighting techniques. We then for mulate our global differentially private test statistic by using the one-shot method, the Gaussian mechanism, and encryption and decryption procedures. Our proposed approach offers several key advantages: By operating under the untrusted server setting, our method resolves the critical weaknesses endemic to trust-based architectures, namely, systemic fragility and exposure to adversarial or governmental targeting. Moreover, the approach incorporates a bounded encryption procedure to ensure secure communication, eliminate the risk of the server actively leaking data, enable easy computation of the estimator's sen sitivity. and circumvent the computation of intractable sensitivities of complex statistics. Furthermore, our proposed method is capable of handling highly correlated covariates and preserving high power for identifying dense but weak signals. Unlike conventional methods, it also avoids the need to handle the Hessian and precision matrices. Simulation studies are carried out to examine the finite-sample behaviour of the proposed method. An application to an adult income dataset is provided.

A Double-Semiparametric Approach for Extending Mixture Cure Models with Interval-Censored Data

Xiaoyu Liu

Jinan University

Mixture cure models (MCMs) have become a valuable tool for analyzing failure time data in settings where a subset of individuals is considered to be "cured" and will never experience failure. Under the MCM framework, the latency component typically postulates a semiparametric survival model for the failure times of uncured individuals, while the incidence component models the probability of being uncured using logistic regression. However, in many practical applications, assuming linear covariate effects or imposing a fully parametric form on the incidence may be unrealistic, given that cure status is inherently unobserved for right-censored individuals. We introduce a more general mixture cure model for interval-censored data by incorporating nonparametric covariate effects in the incidence component. Our method allows semiparametric frameworks for both components of the mixture cure model, offering flexibility in capturing complex relationships between susceptibility and risk factors while preserving interpretability. We develop a spline-based sieve maximum likelihood estimator for both the model parameters and the unknown func-tions, and establish its desirable asymptotic properties. The finite-sample performance and practical utility of the proposed method are demonstrated through a simulation study and an analysis of data from a cardiac allograft vasculopathy study, respectively.

Dually Robust Nonparametric Censored Quantile Regression with Deep Neural Networks

◆ *Ziqiang Deng, Huiqiong Li, Niansheng Tang, Liuquan Sun*

Yunnan University, Yunnan University, Yunnan University, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Censored quantile regression is a flexible tool for survival analysis. However, classical robust methods typically rely on parametric assumption, whereas emerging deep learning alternatives often lack theoretical guarantees and remain

vulnerable to covariate outliers. To address this issue, we propose a neural censored quantile regression method that achieves robustness against both covariate and response outliers. Our approach integrates a novel tempered inverse censoring probability weighting scheme for covariate robustness with a truncated check loss for response robustness. Deep neural networks are employed to approximate the nonparametric conditional quantile function. We establish non-asymptotic error bounds with rates that adapt to the tail behavior of the covariate distribution, recovering the minimax optimal rate when the moment condition strengthens. In addition, we demonstrate that our estimator can mitigate the curse of dimensionality when covariates lie on a low-dimensional manifold. The finite sample performance of the proposed method is examined via extensive simulation studies. An application to a magnetic resonance imaging dataset is provided.

Decoding AI Precision Medicine Paradox: A Fairness-Aware Optimal Transport

Zhimiao Cao

Shandong University

Precision oncology faces a critical paradox where the pursuit of maximal individual-level precision may inadvertently widen healthcare disparities at the group level. This tension necessitates ethically grounded frameworks that reconcile individualized precision with therapeutic equity. In real world oncology, the inherent discreteness of clinical follow-up results in unobserved event times where the exact timing of disease progression remains unknown. This temporal uncertainty compromises the estimation of individualized treatment rules and often leads to suboptimal therapeutic guidance. To bridge this gap, we propose a robust framework that identifies optimal treatment regimes by maximizing the cumulative expected log survival probability. This optimization objective utilizes the logarithmic survival function as the primary metric to ensure the pursuit of maximal individual survival benefits despite the presence of censored clinical outcomes. To maintain therapeutic equity, we incorporate optimal transport to enforce demographic parity by aligning treatment distributions across patient subgroups. This mechanism mitigates treatment disparities tied to sensitive attributes such as gender or ethnicity without sacrificing individual survival benefits. The resulting methodology provides clinicians with a mathematically rigorous tool for equitable precision medicine. By integrating the proposed maximization of individual survival benefits with the mandate for clinical fairness, this work offers a scalable solution for delivering reliable individualized care in practical oncology settings.

S060: Interval-Valued Data Modelling and Forecasting

Least Hausdorff Distance Regression for Interval-Valued Data

◆ *Wei Lin, Xiaojun Song*

University of International Business and Economics, Peking University

The Hausdorff distance is a widely used metric to measure the distance between sets. For interval-valued data, we develop a new regression model based on the Hausdorff distance between two sets which helps to establish useful equivalence relationships between the interval distance and the point distances for associated endpoints, middle points and radii. Correspondingly we propose the Least Hausdorff Distance (LHD) estimator that minimizes the sum of Hausdorff distances between the true and predicted intervals to estimate the parameters of the model. We also provide a proof of consistency and asymptotic normality for the LHD estimator under the assumptions that the interval-valued data are strictly stationary and weakly dependent with alpha-mixing conditions. Monte Carlo simulations show good finite sample properties of the proposed estimators.

Self-Exciting Hysteretic Interval Autoregressive Processes

Han Li, Xinyu Bi, ◆ *Kai Yang, Shishun Zhao*

Changchun University, Changchun University, Changchun University of Technology, Jilin University

Interval time series (ITS) is widely used in fields such as finance and econometrics, due to the rich information contained within the data itself. To model the hysteretic or buffered characteristics of ITS data, this paper introduces a class of self-exciting hysteretic interval autoregressive processes. The key feature of the proposed model is that the regime indicator depends on the threshold variable infinity far away, so that it can effectively explain the hysteretic phenomena in economic. The minimum distance estimation method is develop to address the parameter estimation problem. The nonlinearity test for a linear interval model against a hysteretic interval model is proposed. The forecasting problem is also well solved. Finally, some numerical simulation results and a practical data example are given.

Optimal Parameter-Transfer Learning for Interval-Valued Data

◆ *Haowen Bao, Yongmiao Hong, Yuying Sun, Shouyang Wang*

Xidian University, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

This paper proposes a parameter-transfer learning method for interval-valued time series (ITS) through an optimal model averaging procedure. We develop an innovative approach where a target model incorporates information from auxiliary

source models through shared parametric components, while accommodating distinct structural patterns in interval-valued observations across different model specifications. For optimal weight selection, we propose a novel multifold forward-validation criterion designed for ITS. Under regularity conditions, we establish that the proposed weight estimates achieve asymptotic optimality under misspecified target model, and automatically exclude all non-informative source models when the target model is correctly specified. Monte Carlo simulations demonstrate the favorable finite sample properties of our method. An empirical application to asset price forecasting reveals the superior performance of our approach relative to existing methods. Our findings highlight the potential of machine learning techniques in interval-valued data analysis, offering new insights for financial forecasting under uncertainty.

Autoregressive Models for Functional Interval-Valued Time Series Data

Jianhao Xu

University of Chinese Academy of Sciences

This paper develops autoregressive methods for functional interval-valued time series. Each observation is modeled as an interval-valued random set over a compact domain, and prediction and estimation are based on the support-function D_K distance rather than on an a priori midpoint--range reduction. The proposed FIAR model imposes a common functional lag operator on the two boundary functions, yielding a tractable dynamic restriction that preserves the interval-valued interpretation. We estimate the operator by a D_K-aligned finite-dimensional reduction, a reduced minimum-distance autoregression, and a lifting step back to the functional interval space. For increasing sieve dimension, we establish Hilbert--Schmidt consistency of the lifted operator and isolate the roles of reduced coefficient estimation, empirical eigenspace perturbation, projection bias, and feasible two-step reweighting. Monte Carlo experiments document operator recovery, truncation selection, and feasible reweighting in correctly specified designs. In six illustrative retained financial and climate interval-curve panels, the feasible two-step FIAR forecast delivers the lowest out-of-sample D_K mean distance error among the maintained forecasting rules, with the evidence interpreted as predictive rather than structural.

S066: Modern Methods for Omics Inference and Clinical Translation

Unrolling Regularized Non-Negative Factorization for Spatial Transcriptomics Deconvolution

◆ *Weihong Guo, Liangliang Zhang, Ruitao Liu*

Case Western Reserve University, Case Western Reserve University, University of Utah

Spatial transcriptomics (ST) is revolutionizing biology and medicine by linking gene expression to tissue architecture. This integration provides the spatial resolution and biological context necessary to unravel complex processes such as development, disease progression, immune responses, and therapeutic outcomes. Most current ST technologies (e.g., 10x Visium) measure gene expression at the spot level, where each spot aggregates transcripts from multiple cells. To resolve the underlying cellular composition, reference-guided ST deconvolution methods typically leverages single-cell RNA sequencing (scRNA-seq) data. Starting from a regularized least-squares model integrating graph Laplacian smoothness, simplex constraints, and group sparsity, we derive iterative update rules that form the foundation of our neural network. Each unrolled layer corresponds to one solver iteration. By training end-to-end with synthetic or real data, the model learns all hyperparameters directly, achieving efficient, interpretable, and data-driven inference of cell-type proportions. Ablation study and comparison with some state-of-the-art methods demonstrate the advantages of the proposed method.

CoReg: Co-Expression Network Multivariate Regression for Improved Replicability

◆ *Shuo Chen, Hwiyoung Lee*

University of Maryland, University of Maryland

Accounting for dependence among high-dimensional variables in omics data analysis is critical to obtain accurate and reliable statistical inference. Although latent, omics variables often exhibit structured correlation/co-expression patterns. However, there are few methods explicitly accounting for such structured dependence in the statistical analysis of omics data (e.g., differential expression analysis). To address this methodological gap, we propose a \textbf{Co}expression network multivariate \textbf{Reg}ression (CoReg), which integrates co-expression network structure into multivariate regression analysis to precisely account for the inter-correlations (dependence) among omics variables. We show in simulations that CoReg substantially improves the accuracy of statistical inference and replicability across studies. These findings suggest that CoReg provides an alternative approach for omics data association analysis with dependence adjustment, analogous to the role of mixed-effects models in handling repeated measures in lower-dimensional settings.

DISAA: Differential Identification of Sequence Abundance with Absence for Omics Analysis

Liangliang Zhang

Case Western Reserve University

Microbiome dysbiosis plays a crucial role in human health, and identifying reliable microbial biomarkers is central to understanding disease mechanisms and therapeutic targets. However, microbiome sequencing data are often sparse and zero-inflated, posing major challenges for traditional differential abundance analysis. While DESeq2 remains a gold-standard tool due to its strengths in outlier detection, effect size moderation, overdispersion shrinkage, and multiple testing correction, it performs suboptimally on microbiome data because of inadequate zero handling during normalization and size factor estimation, often leading to inflated false positives. To address this limitation, we introduce DISAA (Differential Identification of Sequence Abundance with Absence), a novel framework that extends DESeq2 through a zero-inflated modeling approach specifically designed for microbiome and other omics data. DISAA achieves a practical balance between model complexity, statistical rigor, and computational efficiency. Simulation studies and real microbiome applications demonstrate that DISAA effectively mitigates zero-related biases, enhances differential detection accuracy, and preserves reproducibility across diverse datasets. This work establishes DISAA as a practical and reliable tool for modern omics research, particularly suited for the sparse and compositional nature of microbiome data.

Bridging Data Gaps in Oncology: Large Language Models and Recommendations for Cancer Treatment in Clinical Trial Applications

Tengjie Tang, Angkai Li, Xuanyu Bai, Xingye Tan, Qingli Ji, Michael Ulis, Lu Si, ◆ *Le Bao*

Pennsylvania State University

The rapid growth of oncology research has generated an enormous volume of data. However, the information is often fragmented across datasets, registries, and publications, limiting their full utility in evidence synthesis and precision oncology. Recent advances in large language models (LLMs) and recommendation algorithms provide opportunities to comprehensively utilize information from multiple data sources across the cancer spectrum and enhance treatment decisions.

We used LLMs to systematically extract and standardize more than 100,000 cancer trials from ClinicalTrials.gov. Each trial was annotated using a customized scoring system reflecting cancer-treatment interactions based on clinical outcomes and trial attributes. Using this structured data set, we implemented 4 state-of-the-art collaborative filtering and hybrid recommendation systems to recommend potentially effective treatments across different cancer types.

The LLM-driven data extraction process successfully generated a comprehensive and rigorously curated database from fragmented clinical trial information with high accuracy, covering 55 cancer types and 12,011 interventions, involving 4,443 distinct drugs. Recommendation models demonstrated high predictive performance (cross-validated RMSE: 0.433--0.698) and identified clinically meaningful new treatments for melanoma, independently validated by oncology experts. We also provide a Shiny app, Clinical Trial Recommendation, which offers an interactive interface for these functionalities.

Our study establishes a proof of concept demonstrating that the combination of LLMs with sophisticated recommendation algorithms can systematically identify novel and clinically plausible cancer treatments. This integrated approach may accelerate the identification of effective therapies for cancer patients, ultimately improving patient outcomes by generating evidence-based treatment recommendations where traditional data sources remain limited.

S068: Modern Perspectives on High-Dimensional Statistical Theory and Methodology

Generalized Tensor Completion with Non-Random Missingness

Jingfei Zhang

Emory University

Tensor completion plays a crucial role in applications such as recommender systems and medical imaging, where data are often highly incomplete. While extensive prior work has addressed tensor completion with data missingness, most assume that each entry of the tensor is available independently with probability p. However, real-world tensor data often exhibit missing-not-at-random (MNAR) patterns, where the probability of missingness depends on the underlying tensor values. This paper introduces a generalized tensor completion framework for noisy data with MNAR, where the observation probability is modeled as a function of underlying tensor values. For model estimation, we develop an alternating maximization algorithm and derive non-asymptotic error bounds for the estimator at each iteration, under considerably relaxed conditions on the observation probabilities. Additionally, we propose a statistical inference procedure to test whether observation probabilities depend on underlying tensor values, offering a formal assessment of the missingness assumption within our modeling framework.

Facilitating Model-Based Clustering Analysis by Dimension Reduction

Yin Jin, ◆ *Wei Luo*

Zhejiang University, Zhejiang University

The Gaussian mixture model (GMM) is commonly used for clustering analysis, but fitting GMM is computationally challenging when the data are large dimensional. To address this issue, we propose a fast dimension reduction procedure that preserves the clustering pattern of the data before applying GMM. The procedure is also applicable for more general unsupervised learning problems centered on detecting non-Gaussian signals from Gaussian noise. We build the general theory for the new type of unsupervised dimension reduction, starting afresh from the identifiable parametrization, and elaborating its applicability in the scenario of GMM. Inspired by Stein's lemma, we propose the first family of dimension reduction methods that are consistent in high-dimensional settings. Using data augmentation, we propose another more general family of methods in connection to the literature of supervised sufficient dimension reduction. The computational efficiency of these methods and their effectiveness for GMM fitting are illustrated numerically at the end.

MintCNA: A Deep Learning Method for Integrative Copy Number Profiling with Single-Cell Multi-Omics Data

Wenhan Bao, ◆ *Feifei Xiao*

University of Florida, University of Florida

Copy number alterations (CNAs) are key drivers of tumor evolution, therapeutic resistance, and disease progression, but accurately resolving CNA landscapes from single-cell data remains challenging. Most existing tools analyze a single omics layer and are vulnerable to assay-specific noises, limiting their ability to recover shared or modality-specific CNAs that are critical for characterizing tumor clonal structure. Existing single-cell multi-omics techniques enable joint sequencing of multiple molecular layers in the same cells, yet in silico methods that fully exploit such paired data for CNA analysis are still missing. Here we present a single-cell multi-omics integration framework, MintCNA, a unified deep learning framework for CNA detection from paired multi-omics data. MintCNA integrates traditional statistical modeling with embedded deep learning structure to enhance CNA profiling across multi-omics. We use an attention-based convolutional autoencoder for data denoising and performs multivariate change-point detection utilizing a local sliding-window screening and ranking procedure. Across various simulations and applications to a colorectal cancer dataset, MintCNA consistently outperforms existing single-omics CNA callers in detection accuracy. MintCNA provides a first single-cell multi-omics integration-based CNA analysis tool to study comprehensive diseases.

Semi-Supervised Linear Regression: Enhancing Efficiency and Robustness in High Dimensions

Kai Chen, ◆ *Yuqian Zhang*

Renmin University of China, Renmin University of China

In semi-supervised learning, the prevailing understanding suggests that observing additional unlabeled samples improves estimation accuracy for linear parameters only in the case of model misspecification. In this work, we challenge such a claim and show that additional unlabeled samples are beneficial in high-dimensional settings. Initially focusing on a dense scenario, we introduce robust semi-supervised estimators for the regression coefficient without relying on sparse structures in the population slope. Even when the true underlying model is linear, we show that leveraging information from large-scale unlabeled data helps reduce estimation bias, thereby improving both estimation accuracy and inference robustness. Moreover, we propose semi-supervised methods with further enhanced efficiency in scenarios with a sparse linear slope. The performance of the proposed methods is demonstrated through extensive numerical studies.

S071: Modern Statistical Learning for Complex Data Analysis

Active Hypothesis Testing under Computational Budgets with Applications to GWAS and LLM

Yin Xia

Fudan University

In large-scale hypothesis testing, computing exact p-values or e-values is often resource-intensive, creating a need for budget-aware inferential methods. This talk proposes 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 eciency under fixed resource limits.

Logistic Regression and Classification with Covariates in Metric Spaces

Yinan Lin, ◆ *Zhenhua Lin*

Chongqing Normal University, National University of Singapore

Inspired by logistic regression, we introduce a regression model for data tuples consisting of a binary response and a set of covariates residing in a metric space without vector structures. Based on the proposed model we also develop a binary classifier for metric-space valued data. We propose a maximum likelihood estimator for the metric-space valued regression coefficient in the model, and provide upper bounds on the estimation error under various metric entropy conditions that quantify complexity of the underlying metric space. Matching lower bounds are derived for the important metric spaces commonly seen in statistics, establishing optimality of the proposed estimator in such spaces. Similarly, an upper bound on the excess risk of the developed classifier is provided for general metric spaces. A finer upper bound and a matching lower bound, and thus optimality of the proposed classifier, are established for Riemannian manifolds. To the best of our knowledge, the proposed regression model and the above minimax bounds are the first of their kind for analyzing a binary response with covariates residing in general metric spaces. We also investigate the numerical performance of the proposed estimator and classifier via simulation studies, and illustrate their practical merits via an application to task-related fMRI data.

EnsLoss: Stochastic Calibrated Loss Ensembles for Preventing Overfitting in Classification

Ben Dai

The Chinese University of Hong Kong

Empirical risk minimization (ERM) with a computationally feasible surrogate loss is a widely accepted approach for classification. Notably, the convexity and calibration (CC) properties of a loss function ensure consistency of ERM in maximizing accuracy, thereby offering a wide range of options for surrogate losses. In this article, we propose a novel ensemble method, namely EnsLoss, which extends the ensemble learning concept to combine loss functions within the ERM framework. A key feature of our method is the consideration on preserving the "legitimacy" of the combined losses, i.e., ensuring the CC properties. Specifically, we first transform the CC conditions of losses into loss-derivatives, thereby bypassing the need for explicit loss functions and directly generating calibrated loss-derivatives. Therefore, inspired by Dropout, EnsLoss enables loss ensembles through one training process with doubly stochastic gradient descent (i.e., random batch samples and random calibrated loss-derivatives). We theoretically establish the statistical consistency of our approach and provide insights into its benefits. The numerical effectiveness of EnsLoss compared to fixed loss methods is demonstrated through experiments on a broad range of 14 OpenML tabular datasets and 46 image datasets with various deep learning architectures. Python repository and source code are available on GitHub.

Weighted Youden Index Maximization

Jialiang Li

National University of Singapore

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.

S084: New Methods for the Analysis of Complex Data

Variational Bayesian Estimation for Joint Models of Longitudinal and Interval-Censored Fairure Time Data with Random Change-Points

Fengting Yi

Yunnan University

In clinical follow-up studies, survival outcomes are often interval-censored, and covariate effects may exhibit abrupt changes across latent thresholds, reflecting heterogeneous disease progression. These features pose substantial challenges for statistical inference due to the intractable likelihood induced by interval censoring and the nonregularity introduced by random change-points. We develop a variational Bayesian approach for estimating random change-points in covariate effects within a joint modeling framework for longitudinal and survival data. The proposed method accommodates interval-censored event times and captures subject-specific change-points in the survival component while accounting for dependence through shared random effects. To address the resulting non-conjugacy, we consider both a non-conjugate variational Laplace approximation and a reparameterization-based gradient estimator for efficient posterior approximation. Simulation studies demonstrate that the proposed methods yield accurate estimation under a range of sample sizes and censoring levels, with the gradient-based approach offering notable computational advantages. An application to Alzheimer's disease data illustrates the ability to identify individualized change-points in the effect of APOE4 on disease progression.

A Longitudinal Multi-Domain Latent Structure Model for Domain-specific Disease Progression

Zijian Ye,Yuanjia Wang,Yu Gu, ◆ *Kai Kang*

Sun Yat-sen University,Columbia University,The University of Hong Kong,Sun Yat-sen University

Parkinson's disease (PD) exhibits substantial heterogeneity, with patients often progressing at different times across motor, cognitive, and non-motor domains. Conventional progression models are based on overall diagnoses or aggregate composite scores, which obscure asynchronous domain-specific progression and limit their usefulness for individualized prognosis and intervention. In this work, we propose a longitudinal multi-domain latent structure model to infer when clinically meaningful progression occurs within each disease domain and how risk factors influence these unobserved domain-specific progression times. The model partitions PD-related markers into clinically distinct domains and characterizes progression within each domain through domain-specific latent disease states. Progression times are treated as fully latent interval-censored events occurring between clinical visits and are inferred from longitudinal marker trajectories with distinct pre- and post-progression distributions. Covariate effects on domain-specific progression times are examined through a Cox proportional hazards model. We establish theoretical guarantees on recovering the marker-domain partition structure. We apply our methods to the Parkinson's Progression Markers Initiative (PPMI) study and identify five heterogeneous clinical domains, each reflecting a specific symptom cluster. The analysis captures individualized, domain-specific progression trajectories and identifies significant associations with gender, education level and putamen area atrophy.

Transfer Learning Estimation of the Accelerated Failure Time Model Based on High-Dimensional Data

◆ *Yichen Lou, Mingyue Du, Hui Zhao, Jianguo Sun*

Nanyang Technological University, Singapore, Jilin University, Zhongnan University of Economics and Law, Southern University of Science and Technology

The accelerated failure time model is one of the most commonly used models for regression analysis of failure time data and many methods have been developed for its estimation. However, standard approaches may fail or underperform when available information is limited. To address this issue, we propose two transfer learning estimation procedures that leverage auxiliary information from multiple source datasets. The first is a data-driven source detection procedure that classifies the source datasets into positively and negatively transferable groups and performs estimation using only the positively transferable or informative source datasets. The other is an ensemble- based approach that adaptively assigns weights to source datasets based on their relevance to the target dataset. Theoretical justifications are provided for the proposed methods, and an extensive simulation study is performed and indicates that the proposed methods work well in practice. Finally, they are applied to a set of real data on the end-stage renal disease that motivated this study.

S085: New Statistical and Coputational Methods for Large Data Analysis

Causal Learning with Invalid Instruments for High-dimensional Imaging Responses

Yiting Wang, Chunlin Li, ◆ *Shan Yu, Lily Wang*

University of Virginia, University of Virginia, University of Virginia, George Mason University

Large-scale neuroimaging studies provide high-dimensional imaging phenotypes for understanding brain-related diseases and their causal relations with risk factors. While instrumental variable (IV) estimation is effective for uncovering causality, existing IV methods face significant limitations when applied to high-dimensional imaging data, particularly in handling large-scale data and incorporating the intrinsic spatial structure of images. To address these challenges, we propose a novel framework, High-dimensional Functional-imaging Instrumental Variable Estimation (Hi-FIVE), which enables estimation of spatially varying causal effects in the presence of many invalid IVs. Our approach uses nonparametric spatial smoothing through multivariate penalized splines to capture the spatial structure of imaging data, combined with distributed-learning-based instrumental variable selection to handle potentially invalid instruments and large-scale data. We establish rigorous theoretical guarantees for the Hi-FIVE framework, including new identifiability results and the asymptotic properties of the estimator. Building on these theoretical results, we further construct data-driven simultaneous confidence regions for statistical inference. Hi-FIVE is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) to illustrate its effectiveness in uncovering causal effects in large-scale observational neuroimaging studies.

High-Dimensional Dynamic Covariance Models with Random Forests

Shuguang Yu, Fan zhou, Yingjie zhang, ◆ *ziqi chen, hongtu zhu*

East China Normal University

This paper introduces a novel nonparametric method for estimating high-dimensional dynamic covariance matrices with multiple conditioning covariates, leveraging random forests and supported by robust theoretical guarantees. Unlike traditional static methods, our dynamic nonparametric covariance models

effectively capture distributional heterogeneity. Furthermore, unlike kernel-smoothing methods, which are typically limited to conditioning covariates with one or two dimensions, our approach accommodates covariates of dimensionality greater than two within a fully nonparametric framework. To the best of our knowledge, this is the first method to use random forests for estimating high-dimensional dynamic covariance matrices. In high-dimensional settings, we establish uniform consistency theory, providing nonasymptotic error rates and model selection properties, even when the response dimension grows sub-exponentially with the sample size. These results hold uniformly across a range of conditioning variables. The method's effectiveness is demonstrated through simulations and a stock dataset analysis, highlighting its ability to model complex dynamics in high-dimensional scenarios.

Detection of Medication Taking Using Wrist-Worn Commercially Available Wearable Device

◆ *Quy Cao, Amy Laughlin*

University of Pennsylvania, University of Pennsylvania

Medication nonadherence is a persistent and costly problem across health care. Measures of medication adherence are ineffective. Methods such as self-report, prescription claims data, or smart pill bottles have been used to monitor medication adherence, but these are subject to recall bias, lack real-time feedback, and are often expensive. We proposed a method for monitoring medication adherence using a commercially available wearable device. Passively collected motion data were analyzed on the basis of the Movelet algorithm, a dictionary learning framework that builds person-specific chapters of movements from short frames of elemental activities within the movements. We adapted and extended the Movelet method to construct a within-patient prediction model that identifies medication-taking behaviors.

Using 15 activity features recorded from wrist-worn wearable devices of 10 patients with breast cancer on endocrine therapy, we demonstrated that medication-taking behavior can be predicted in a controlled clinical environment with a median accuracy of 85%. These results in a patient-specific population are exemplar of the potential to measure real-time medication adherence using a wrist-worn commercially available wearable device.

Multiple Change Point Detection in Directed Temporal Networks with Non-Gaussian Components

◆ *Sehun An, Ting Fung Ma, Shanghong Xie*

University of South Carolina., University of South Carolina, University of South Carolina

Biomarker networks, such as brain networks, often exhibit dynamic and nonstationary behavior. In this work, we propose a framework for modeling nonstationary directed temporal networks with multiple subjects by piecewise stationary models with unknown number of change points for stages of brain activities. A semi-parametric model with a non-Gaussian component is used for robust inference and is likelihood-free. Based on the minimum description length (MDL) principle, we simultaneously estimate the number and locations of change points along with the directed network within each segment via an optimization problem. We establish theoretical guarantees, such as consistency and asymptotic distributions of model parameters, including the number of change points, change point locations, and network parameters under mild conditions. We show that the convergence rates of estimated change point locations and network parameters depend on the number of subjects and time points, respectively.

Computational scalability of the procedure is ensured by a pruned dynamic programming algorithm. Extensive simulation studies are presented to assess the finite sample performance of the proposed procedure. We apply the proposed method to ADHD fMRI data to characterize time-varying directed functional connectivity and identify distinct connectivity states.

S096: Recent Advances in Computational Methods for Single-Cell and Spatial Omics

Intelligent Decoding and Generation of Spatial Biology

Shihua Zhang

The Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Technological advances in spatial transcriptomics are critical for better understanding the structures and functions of tissues in biological research. The combination of intelligent or statistical algorithms and spatial transcriptomics has emerged to pave the way for deciphering tissue architecture. We have made great efforts to advance intelligent spatial transcriptomics and developed a group of STA- tools such as STAGATE, STAligner, STAMarker, STAGE, STASCAN, STALocaor, and STAMapper. In this talk, I will take these tools as examples to introduce how to utilize deep learning to model spatial transcriptomics data effectively. Moreover, I will shed light on some recent progress in spatial multi-omics integration with STAMO, directionally or temporally variable gene detection with STAVAG, and high-fidelity simulation and full-view 3D modeling of spatial transcriptomics with STADiffuser.

Intelligent Analysis and Mining of Spatial Omics Data

Zhiyuan Yuan

Fudan University

Cell fate is jointly determined by intracellular molecular regulatory networks and the extracellular microenvironment. Single-cell omics technologies can measure intracellular molecular expression profiles but lose spatial location information. In recent years, the rapid development of spatial omics technologies has enabled researchers to acquire spatial location information in addition to measuring rich molecular expression profiles, providing conditions for deciphering tissue microenvironments. This presentation will introduce our team's computational methodological research in the field of spatial omics.

Predicting Unseen Perturbation Effects Using Data Similarity and Global Structure

Jingshu Wang

University of Chicago

Predicting cellular responses to genetic perturbations is a central challenge in modern computational biology, with important implications for understanding gene function and designing therapeutic interventions. Using large-scale genome-wide Perturb-seq data, we show that perturbation effects exhibit substantial similarity across cell lines. Surprisingly, simple averaging of effects from other cell types consistently outperforms learned models, especially for the strongest gene expression responses. In contrast, low-rank embedding approaches perform worse than expected, indicating that biologically relevant effects are not well captured by low-dimensional representations.

We further find that the rejection matrix is dense and approximately rank-1, reflecting a broadly shared pattern of weak effects across perturbations and genes. As weaker effects are filtered out, this global structure diminishes and more localized, biologically meaningful modules emerge. Together, these results suggest a high-dimensional perturbation landscape consisting of dense, diffuse weak effects alongside sparser, perturbation-specific strong effects that are shared across cell types but not low-dimensional.

Spatial-Aware Detection of Copy Number Alterations from Spatial Transcriptomics Using SpaCNA

Zihui Zhang, Xiaochen Wang, Hong Xuan, Yan Xu, ◆ Zijie Jin, Ruibin Xi

Peking University, Peking University, Peking University, Chinese Academy of Medical Sciences & Peking Union Medical College, Peking University, Peking University

Spatial transcriptomics (ST) profiles genome-wide gene expression while preserving spatial context, yet accurate detection of copy number alterations (CNAs) in tumor ST data remains challenging. Here we present SpaCNA, a computational framework that integrates multi-modal information of ST for robust CNA detection. SpaCNA aggregates expression from neighboring spots with similar morphological features and leverages a hidden Markov random field model incorporating spatial continuity for reliable CNA detection in ST datasets. Further, SpaCNA can reconstruct 3D CNA profiles with spatial continuity across consecutive slices when applied to 3D ST datasets. Extensive benchmarking on simulated data and real cancer datasets demonstrates superior accuracy, achieving up to 0.95 F1-score in CNA detection and tumor region identification. In real-world applications to breast cancer and colorectal cancer, SpaCNA reveals tumor boundaries and spatially distinct subclones with context-dependent interactions within the microenvironment. Notably, SpaCNA performs CNA detection in a 3D ST dataset of head and neck squamous cell carcinoma, revealing the tumor evolution trajectory of three subclones in 3D space. By providing accurate CNA inference, SpaCNA facilitates the analysis of intratumoral heterogeneity and spatial cancer biology.

S103: Recent Advances in Network Analysis

Differentially Private Estimation and Inference for Spatial Autoregressive Models

Ziyi Kong, Yuxin Tao, ◆ Danyang Huang, Zhanrui Cai

Renmin University of China

Privacy-preserving data analysis has attracted increasing attention in modern statistics. However, how to simultaneously protect both the network structure and the nodal information remains a challenging problem. This paper investigates differentially private estimation and inference for Spatial Autoregressive (SAR) models, aiming to protect the privacy of network structure and individual nodal information. We first derive the minimax lower bounds for differentially private estimators under the SAR model, providing a theoretical benchmark for optimal algorithm design. Building on these insights, we propose a differentially private algorithm for SAR parameter estimation and present a comprehensive analysis of its convergence properties, including examinations of representative cases in specific networks. Additionally, we develop differentially private inference procedures for the SAR model. Empirical validation through simulations and real-world data analysis confirms the reliable finite-sample performance of the proposed methodology.

Recover Common Components in Heterogeneous Chain-Linked Matrices

◆ *Xinyan Fan, Hao Mei, Shuangge Ma*

Renmin University of China, Renmin University of China, Yale School of Public Health

We consider the integrated analysis of relation matrices from multiple data sources, where existing methods are inadequate for settings in which cross-source heterogeneity and partially overlapping node sets are present. We develop a novel method based on aggregated projection matrices constructed from the spectral embeddings of overlapping individuals. Leveraging projection techniques and spectral decomposition, we propose an effective algorithm for extracting the common component matrix shared across sources. Advancing from literature, we allow the number of data sources to diverge, establish non-asymptotic error bounds for estimating the common component, and propose a theoretically consistent procedure for selecting its rank. The analysis scope is then expanded to accommodate covariate-specific effects. Simulation demonstrates effectiveness of the proposed approach. We analyze disease co-occurrence with the Taiwan Health Insurance Research Database, which demonstrates practical utilization of the proposed approach and leads to sensible findings.

Graph-Based Multisample Comparison with Application to Feature Selection for Multi-Category Responses

◆ *Dan Pu, Haomin Shi, Wei Lan, Chih-Ling Tsai*

Southwestern University of Finance and Economics

This article proposes a graph-based maximum pairwise difference (MPD) test to compare K-sample distributions. This test can be utilized to identify the most informative variables contributing to the heterogeneity of multiple multivariate distributions. To construct the test, we first pool observations from multiple samples to create a similarity graph based on the distances between any two observations. For each observation, we next evaluate its connection probability to each sample. Then, we doubly maximize the squared differences of these connection probabilities via the within and between samples. Accordingly, the MPD test is a max-of-squares test and can effectively detect sparse alternatives. In addition, we show the asymptotic null distribution of the test statistic and establish its consistency under alternatives. To illustrate the finite sample performance of the proposed test, we conduct extensive simulation studies. The results indicate that the MPD test outperforms some commonly used multisample tests. Finally, we employ it to select relevant features related to corporate credit ratings.

S118: Recent Progress in Dependent Data Analysis

Hidden-Factor Modeling of Climate-Driven Losses: An Illustration Using U.S. Crop Indemnities

◆ *Kun Chen, Xingzuo He, Rui Huang, Zhiwei Tong*

Southwestern University of Finance and Economics, Southwestern University of Finance and Economics, Nanjing University of Chinese Medicine, University of Iowa

Climate risk induces complex dependence across losses at geographically distant locations. Accurately forecasting the joint distribution of losses is therefore important for portfolio-level risk management, yet remains challenging and relatively underexplored. Using U.S. crop insurance indemnities as a representative setting, this study argues that dependence in climate-driven losses arises from two distinct sources: local spatial spillovers and latent common climate factors. To jointly capture these sources of dependence, we adopt a spatial panel regression model that incorporates both a spatial autoregressive component and a latent factor component. Using data from multiple representative counties across five Farm Resource Regions, we generate one-year-ahead predictive distributions at the county level and subsequently the predictive distribution of aggregate crop indemnities. We compare the proposed model with baseline models that exclude either the latent factor component or both sources of dependence. The results show that the model ignoring both components substantially underestimates aggregate indemnities, while the model relying solely on spatial spillovers misattributes the effects of latent factors to local spillovers. In contrast, the main model, by allowing for heterogeneous responses to common climate regimes, hence, the possibility of capturing negative dependence, achieves markedly improved in-sample fit, more accurate out-of-sample median forecasts that closely track realized aggregate indemnities, and substantially tighter prediction intervals without loss of coverage. These improvements translate into a more economically grounded assessment of diversification potential, more balanced capital allocations, and more credible stress-testing outcomes.

Multivariate Contaminated Normal Linear Mixed Models with Censored and Missing Outcomes Applied to Alzheimer's Disease Data

◆ *Wan-Lun Wang, Tsung-I Lin*

Cheng Kung University, Chung Hsing University

This paper proposes a robust approach to jointly modeling multiple repeated clinical measures with intricate features. More specifically, we aim to expand the scope of the multivariate linear mixed model by using the multivariate contaminated normal distribution. The proposed model, called the MCNLMM-CM, is designed to handle minor outliers effectively while simultaneously accommodating censored measurements and intermittent missing responses. An expectation conditional maximization either (ECME) algorithm is developed to estimate the parameters of

the proposed model in situations involving missing at random responses. We also provide techniques for approximating the asymptotic standard errors of model parameters, recovering censored data, imputing missing values, and identifying outliers. The proposed methodology is inspired by and applied to data from the Alzheimer's Disease Neuroimaging Initiative cohort study, which involves longitudinal clinical measurements of patients with mild cognitive impairment. A simulation study is conducted to evaluate the finite-sample properties of the parameter estimators.

Scalable Bayesian Spatiotemporal Tensor Modeling for Censored and Missing Areal Data

Luis Mauricio Castro Cepero

Pontificia Universidad Católica de Chile

We propose a new Bayesian approach for spatiotemporal areal data with censored and missing observations, where the response is represented as a tensor indexed by spatial units and time points. The method introduces a flexible random effect that combines the spatial dependence structures of the Simultaneous Autoregressive (SAR) and Directed Acyclic Graph Autoregressive (DAGAR) models with a temporal autoregressive component. This formulation brings both spatial models into a unified spatiotemporal framework by expressing them as Gaussian Markov random fields in innovation form, providing an interpretable representation of spatial, temporal, and joint spatiotemporal dependence while preserving the multiway structure of the data. The innovation-based structure enables scalable Bayesian implementation in \texttt{Stan} for datasets of moderate size. Simulation studies show that the proposed model outperforms common ad hoc imputation strategies for censored and missing data. We further illustrate its practical value through an application to carbon monoxide (CO) concentrations, where the proposed framework captures complex spatiotemporal dependence in the presence of censoring and missingness.

Generative Bootstrap Sampler for Scalable Inference Under Dependent Data

◆ *Fanghua Chen, Ting Fung Ma, Ray Bai*

University of South Carolina, University of South Carolina, George Mason University

Massive datasets with complex and structured dependence, such as spatial and spatio-temporal data, create long-standing challenges for statistical learning and uncertainty quantification. Classical approximation methods, including Vecchia methods and composite likelihood, often become unstable due to growing approximation error, while nonparametric approaches such as the bootstrap require repeated resampling and model refitting, resulting in substantial computational complexity and limited scalability for large dependent datasets.

We introduce Generative Bootstrap Sampler (GBS), a scalable framework that learns the bootstrap operator under general dependent data structures. GBS constructs a generative mapping from random bootstrap weights to implicit sampling distributions. This mapping is learned through a single optimization procedure and can be evaluated at negligible cost once trained. The framework naturally integrates likelihood approximations such as Vecchia methods, enabling efficient inference for large spatial and spatio-temporal processes.

The GBS method provides stable and theoretically grounded uncertainty quantification under dependent observations, achieves substantial computational gains while preserving inferential validity. The approach offers a unified and scalable solution for modern statistical learning problems involving massive dependent data.

S133: Statistical Inference for Complex Data

Is the F-Test Doubly Robust?

◆ *Lucy Xia, Oliver Feng, Yang Feng, Min Xu, Richard Samworth*

The Hong Kong University of Science and Technology, London School of Economics and Political Science, New York University, Rutgers University, University of Cambridge

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.

Efficient Computing for High-Dimensional Portfolio Optimization

Songshan Yang

Renmin University of China

We focus on efficient computing for high dimensional portfolio optimization when the number of assets is large relative to the estimation window. The first part studies an effective-parameter formulation and develops an asset-splitting ADMM algorithm together with a folded-concave refinement and oracle type theory. The second part studies Dantzig-type portfolio optimization, which estimates sparse portfolio weights directly under the budget constraint. I will discuss its Lasso and SCAD versions, the associated parallel algorithms, and empirical results from simulations and U.S. equity data, including the S&P 500 and Russell2000.The comparison highlights the difference between indirect effective-parameter estimation and direct sparse-weight estimation, with particular emphasis on scalable computation.

Conditional Selective Inference for the Selected Groups in Panel Data

◆ *Chuang Wan, Jiajun Sun, Xingbai Xu*

Jinan University, Xiamen University, Xiamen University

This paper addresses the problem of testing differences in group-specific slopes between clusters identified via k-means clustering in panel data. In this setting, classical Wald-type tests are inherently invalid, as they yield severely inflated Type I error probabilities. This failure stems from "double dipping"—using the same dataset for both group identification and subsequent inference. To resolve this, we propose a valid selective inference approach that conditions on the selection event to account for data-driven clustering. We derive an efficient algorithm to compute exact p-values for k-means outcomes. Our framework is notably flexible: it requires only the asymptotic normality of initial estimators and extends naturally to GMM estimation, making it applicable to dynamic panel models and panel data models with endogenous variables. Simulation studies demonstrate robust finite-sample performance. Applying our method to the global relationship between economic growth and CO2 emissions, we uncover novel evidence of cross-country heterogeneity.

Fairness Auditing with General Protected Attribute

Junjie Huang, Jingyuan Liu, ◆ *Xianli Zeng*

Xiamen University

While fair algorithmic design has advanced rapidly, fairness auditing remains underdeveloped, especially for continuous protected attributes such as age or income. Directly applying existing methods to continuous protected attributes may render unfair interval unidentifiable. In this paper, we provide a unified statistical framework for fairness auditing that accommodates both discrete and continuous sensitive values. Our approach builds on nonparametric methods to construct a test statistic that quantifies disparities as a function of protected attributes. We derive the asymptotic distribution of the proposed test statistics, based on which we establish the fairness auditing strategy.

For practical implementation, we further develop a computationally efficient multiplier Bootstrap procedure to circumvent the need for explicit variance estimation. Simulation studies show that the proposed test procedure achieves comparable performance to existing methods in the discrete case while significantly outperforms them when the protected attribute is continuous. We apply our method to the COMPAS datasets, revealing substantial age-related disparities that would be masked by existing approaches.

S141: Statistical Methods for Genetics and Genomics

GenoSim: Feasible Correlations of Discrete Random Variables and Genotype Data Simulation

Hong Zhang, ◆ *Zheyang Wu*

Pfizer, Worcester Polytechnic Institute

Simulating correlated genotype data with realistic linkage disequilibrium (LD) structures is a fundamental problem in statistical genetics. A prerequisite question is: given two discrete random variables with specified marginals (e.g., genotypes with specified genotype frequencies), what are the feasible boundaries for their correlation or other dependence measures? We develop a general linear programming (LP) framework that computes sharp feasibility bounds for a broad class of dependence measures — including Pearson and Spearman correlations, Cohen's Kappa, and mean absolute error — between any two discrete random variables with finite support. This framework substantially extends the classical Frechet-Hoeffding bounds approach by accommodating arbitrary objective functions and additional constraints on joint cell probabilities. Moreover, we develop an efficient NORTA (Normal-To-Anything) simulation procedure for correlated genotype data. We show that genotype-level simulation — modeling variants as discrete random variables directly (e.g., binomial) rather than the traditional haplotype-based approach (which assumes summation of Bernoulli random variables) — covers a strictly larger correlation domain, enabling more realistic LD structures.

Mendelian Randomization Methods for Pleiotropic and Weak Effects and Their Applications

◆ *Miaoxin Li, Bin Tang*

Sun Yat-Sen University, Sun Yat-Sen University

Mendelian randomization (MR) harnesses genetic variants as instrumental variables to infer causal relationships between exposures and outcomes. Its validity depends on the three IV assumptions.

First, the relevance assumption is the most basic and often overlooked assumption in MR, as it can be guaranteed by GWAS. However, an insufficient number of IVs (<3) often arises from limited GWAS sample sizes or low trait heritability, forcing researchers to relax variant selection thresholds (e.g., using more lenient p-value thresholds) to obtain enough IVs. This practice, in turn, violates the relevance assumption and degrades estimation accuracy and statistical power. To address this issue, we develop ICMR (Integrating Credibility for Mendelian Randomization), which incorporates the credibility of IVs being truly associated with the exposure. ICMR overcomes the false-positive challenge induced by threshold relaxation, enabling more effective causal inference and delivering consistent results across various thresholds.

Second, the independence assumption and the exclusion restriction assumption are unverifiable; violating them leads to correlated and uncorrelated horizontal pleiotropy in MR, respectively. To address this issue, we propose PCMR (Pleiotropic Clustering framework for Mendelian Randomization) to model both types of pleiotropy. PCMR detects correlated horizontal pleiotropy and extends the zero modal pleiotropy assumption to enhance causal inference in trait pairs where correlated horizontal pleiotropic variants are present.

Spatially-Informed Reference-Free Deconvolution for Spatial Transcriptomics with SpatialCD

Phuong Vo, ◆ *Yuehua Cui*

Michigan State University, Michigan State University

Cell-type deconvolution has been instrumental for the analysis of spatial transcriptomics (ST) data to reveal underlying tissue heterogeneity. While reference-based methods have been widely explored, practical limitations, particularly the need for matched scRNA-seq datasets, highlight the value of robust reference-free methods. Existing reference-free approaches, such as STdeconvolve, overlook spatial information, despite the well-established observation that spatially adjacent spots often share similar cellular compositions. Motivated by this, we propose SpatialCD, a Spatially-informed reference-free Cell-type Deconvolution method that extends Latent Dirichlet Allocation (LDA) model with spatial regularization to encourage neighboring spots to share similar cell-type compositions. SpatialCD produces improved estimates of cell-type proportions and gene expression profiles. Across simulated and real datasets, including MERFISH-derived simulations, mouse olfactory bulb (MOB), 10x Visium, and DBiT-seq data, SpatialCD consistently outperforms existing reference-free methods (e.g., STdeconvolve and SpiceMix), recovering more accurate transcriptional patterns and revealing biologically coherent spatial organization. This work advances statistical tools for spatial transcriptomics and enriches the methodological toolkit for complex spatial gene expression analysis.

TD-TMA：Integrative Analysis of Multiple Ligand-Receptor Families for Sample Clustering via Tensor Decomposition Based Trimmed Model Averaging

◆ *Nayang Shan, Lin Hou, Xiting Yan, Zuoheng Wang*

Capital University of Economics and Business, Tsinghua University, Yale University, Yale University

While single-cell RNA sequencing provides the resolution necessary to infer intercellular interactions, aggregating these complex interaction profiles into robust patient subtyping remains a significant challenge. Current methods utilizing tensor decomposition show promise in capturing high-order communication patterns but often suffer from high data sparsity and prohibitive computational overhead during the clustering process. Furthermore, standard ensemble strategies typically rely on the indiscriminate integration of base partitions, which allows noise-dominated models to dilute the consensus structure and inherently lacks interpretability. We propose Tensor Decomposition-based Trimmed Model Averaging (TD-TMA), a novel framework designed for robust sample clustering. TD-TMA first constructs 4-dimensional tensors by integrating ligand-receptor mode information, subsequently extracting latent communication factors via tensor decomposition. To enhance robustness and performance, the framework employs a subsampling strategy on the sample factor loadings for similarity network fusion. Crucially, it incorporates a trimming mechanism based on the average silhouette width to exclude noise-dominated, low-fidelity partitions prior to consensus integration. The resulting weighted fused matrix and derived mode importance scores facilitate both accurate sample clustering and biological interpretation. Across three benchmark datasets, TD-TMA consistently outperforms Tensor-cell2cell and standard Similarity Network Fusion, effectively mitigating noise and revealing biologically distinct sample clusters.

S142: Statistical Methods for Reliable Machine Learning

Personalized Bayesian Federated Learning with Wasserstein Barycenter Aggregation

Feng Zhou

Renmin University of China

Personalized Bayesian federated learning (PBFL) handles non-i.i.d. client data and quantifies uncertainty by combining personalization with Bayesian inference. However, existing PBFL methods face two limitations: restrictive parametric assumptions in client posterior inference and naive parameter averaging for server aggregation. To overcome these issues, we propose FedWBA, a novel PBFL method that enhances both local inference and global aggregation. At the client level, we use particle-based variational inference for nonparametric posterior representation. At the server level, we introduce particle-based Wasserstein barycenter aggregation, offering a more geometrically meaningful approach. Theoretically, we provide local and global convergence guarantees for FedWBA. Locally, we prove a KL divergence decrease lower bound per iteration for variational inference convergence. Globally, we show that the Wasserstein barycenter converges to the true parameter as the client data size increases. Empirically, experiments show that FedWBA outperforms baselines in prediction accuracy, uncertainty calibration, and convergence rate, with ablation studies confirming its robustness.

A Unified Newton Framework for Perturbation of Structured Matrix Estimators

Ziliang Zhong, ◆ *Shuyang Ling*

New York University Shanghai, New York University Shanghai

The spiked matrix model plays a central role in high-dimensional probability and statistics and has long been a fundamental object of study in random matrix theory. A key question concerns the distribution of leading eigenvectors. While these questions have been extensively analyzed using probabilistic techniques, we propose a complementary, computationally motivated framework based on Riemannian optimization. Exploiting the variational characterization of leading eigenvectors, we analyze a one-step Riemannian Newton iteration on the Stiefel (or Grassmannian) manifold associated with the spectral optimization problem. This approach yields an explicit approximation of the leading eigenspace together with a sharp second-order expansion that quantifies its deviation from the global optimizer. The analysis is purely algebraic and does not rely on specific distributional assumptions, making the framework applicable to a broad class of noise models.

Beyond the eigenvalue problem, the framework also naturally extends to the characterization of singular values and includes applications to orthogonal synchronization problems. Our results provide an algebraic and computational perspective on eigenvector fluctuations and offer an alternative route toward distributional characterization.

Toward Exact Convergence in Byzantine-Robust Decentralized Learning: A Statistical Identification Approach

Siyuan Zhang, ◆ *Chengde Qian, Xin Liu, Changliang Zou*

Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai Jiao Tong University, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Nankai University

To defend against Byzantine attacks in decentralized learning, most existing methods rely on robust aggregation rules to mitigate the influence of malicious machines. However, these strategies inherently introduce bias, leading to inexact convergence with non-vanishing steady-state errors. In this paper, we propose a strategic shift from passive aggregation to active identification by introducing the Decentralized Rescaled Stochastic Gradient Descent with Byzantine Machine Identification (DRSGD-ByMI) framework. The core of our approach is an identification-based ``detect-then-optimize'' pipeline, where a p-value-free detection procedure is developed to accurately prune malicious nodes from the network. By leveraging sample-splitting score statistics, this identification mechanism achieves false discovery rate control without requiring restrictive distributional assumptions. We theoretically demonstrate that this precise identification allows the decentralized network to recover sufficient connectivity among the normal nodes, thereby enabling DRSGD-ByMI to match, even in the presence of Byzantine machines, the same order-optimal convergence rate as standard decentralized stochastic first-order methods. Numerical experiments validate our theoretical results and demonstrate the effectiveness of DRSGD-ByMI for decentralized robust learning problems.

False Discovery Rate Control over Data-Driven Subgroups: A Unified Conformalized Approach

◆ *Yuyang Huo, Xiaoning Wang, Liuhua Peng, Changliang Zou*

Nankai University, Nankai University, Melbourne university, Nankai University

Multiple testing over subgroups of interest is crucial for many scientific problems and downstream decisions. In cases where prior knowledge of group partition is unavailable, a data-driven subgrouping algorithm is applied to the data. Controlling the false discovery rate (FDR) in such data-driven subgroups is challenging because the subgrouping process distorts the distribution of test statistics, creates complex dependence within groups, and results in a random number of hypotheses. To overcome these issues, we leverage exchangeability concepts in conformal inference and propose a conformalized approach named SCARF (\textbf{S}wapping-\textbf{CA}librating-\textbf{R}e\textbf{F}erencing), which consists of three key components: a \emph{swapping} re-subgrouping procedure to construct selective p-values, a \emph{calibrating} procedure to decouple the within-group dependence, and a \emph{referencing} step to adjust for the randomness in subgroup sizes. Unlike existing methods based

on marginal p-values, our approach constructs selective p-values by incorporating a calibration set, fully utilizing subgrouping information to significantly boost statistical power. We demonstrate that SCARF achieves finite-sample FDR control in a distribution-free and algorithm-agnostic regime. Moreover, we establish the oracle decision rule for subgroup FDR control and introduce a boosted SCARF variant that achieves asymptotic optimality under subgrouping stability. Numerical experiments on both synthetic and real datasets confirm that SCARF not only maintains valid FDR control over data-driven subgroups but also outperforms benchmarks across various subgrouping algorithms and data-generating scenarios.

S143: Statistical Modeling and Inference for Large Language Models

Fundamental Trade-Offs in Multi-Bit Watermarking of Stochastic Processes

◆ *Haiyun He, Yepeng Liu, Zhuoer Shen, Ziqiao Wang, Yongyi Mao, Yuheng Bu*

The Hong Kong University of Science and Technology (Guangzhou), University of California, Santa Barbara, University of California, Santa Barbara, Tongji University, University of Ottawa, University of California, Santa Barbara

We study multi-bit watermarking for data generated by stochastic processes, where a hidden message is embedded during sampling and must be decodable by an authorized detector that possesses side information unavailable to unauthorized observers. In high-stakes deployments, a practical watermark must simultaneously control false alarms, preserve generation quality without distorting the output distribution, and support reliable multi-bit decoding. Satisfying all three goals at once inevitably creates fundamental trade-offs. We formulate watermark embedding as a distributional information-embedding problem and watermark detection as a multiple-hypothesis testing problem under distortion and rate constraints, leading to four fundamental metrics: false-alarm probability, detection error probability, distortion, and information rate. Within this information-theoretic framework, we derive matched converse and achievability bounds that characterize the optimal trade-offs and provide scheme-agnostic benchmarks for any watermarking method. For stationary ergodic stochastic processes, we further obtain matched asymptotic limits and connect them to the finite-sample regime. Finally, we present a reference watermarking construction satisfying our assumptions and empirically illustrating the predicted trade-offs.

Provable Training Data Identification for Large Language Models

Hongxin Wei

Southern University of Science and Technology

Identifying training data of large-scale models is critical for copyright litigation, privacy auditing, and ensuring fair evaluation. However, existing works typically treat this task as an instance-wise identification without controlling the error rate of the identified set, which cannot provide statistically reliable evidence. In this work, we formalize training data identification as a set-level inference problem and propose Provable Training Data Identification (PTDI), a distribution-free approach that enables provable and strict false identification rate control. Specifically, our method computes conformal p-values for each data point using a set of known unseen data and then develops a novel Jackknife-corrected Beta boundary (JKBB) estimator to estimate the training-data proportion of the test set, which allows us to scale these p-values. By applying the Benjamini–Hochberg (BH) procedure to the scaled p-values, we select a subset of data points with provable and strict false identification control. Extensive experiments across various models and datasets demonstrate that PTDI achieves higher power than prior methods while strictly controlling the FIR.

Bilevel ZOFO: Efficient LLM Fine-Tuning and Meta-Training

Peiran Yu

University of Texas Arlington

Fine-tuning pre-trained Large Language Models (LLMs) for downstream tasks using First-Order (FO) optimizers presents significant computational challenges. Parameter-Efficient Fine-Tuning (PEFT) methods address these by freezing most model parameters and training only a small subset. However, PEFT often underperforms compared to full fine-tuning when high task-specific accuracy is required. Zeroth-Order (ZO) methods fine-tune the entire pre-trained model without back-propagation, estimating gradients through forward passes only. While memory-efficient, ZO methods suffer from slow convergence and high sensitivity to prompt selection. We bridge these two worlds with Bilevel-ZOFO, a bilevel optimization method that couples fast, local FO-PEFT adaptation at the inner level with stable, memory-efficient ZO updates of the full backbone at the outer level. The FO-PEFT inner loop performs fast, low-memory local adaptation that reduces the variance of ZO estimates and stabilizes the search, guiding the outer ZO updates of the full backbone and reducing prompt sensitivity. In the mean time, the outer ZO provides better generalization ability for PEFT. We provide theoretical convergence guarantees and empirically demonstrate that Bilevel ZOFO significantly outperforms existing ZO and FO-PEFT methods, achieving 2–4× faster training while maintaining similar memory efficiency. Additionally, we show by updating the backbone with ZO and adapting only a tiny FO-PEFT block per

task, Bilevel-ZOFO combines full-model capacity with few-shot efficiency, making it a very efficient meta-learning algorithm that quickly adapts to new tasks.

Optimal Watermark Localization in Mixed-Source Large Language Model Texts

Xiang Li

University of Pennsylvania

Watermarking provides a principled way to authenticate text generated by large language models (LLMs). In real-world use, however, the final text may be mixed-source: some parts may come from watermarked LLM outputs, while others may be rewritten, inserted, deleted, paraphrased, or otherwise modified. As a result, only part of the final text may still preserve watermark evidence. While prior work has studied global detection of surviving watermark signals, it remains unclear when such evidence can be localized. In this paper, we formulate watermark localization as a token-level multiple testing problem based on pivotal statistics, where each position has a latent indicator denoting whether watermark evidence is present. We characterize the statistical limits of this problem under a regime that captures sparse surviving signals, concentrated token distributions, and growing vocabularies. We establish sharp phase transitions for three inference goals, namely global detection, discovery, and classification, showing in particular that discovery is strictly harder than detection, while exact classification is impossible under the general mixed-source model. We then propose an adaptive thresholding method that localizes watermark evidence without requiring any problem-dependent parameters. We prove that our method achieves the optimal discovery boundary and attains near-optimal discovery power among natural local rules. Experiments on synthetic and LLM-generated texts confirm our theoretical findings and demonstrate the effectiveness of the proposed method under common human edits.

S149: The Broad Roles of Design in Modern Statistical Analysis

From Subsampling to Causal Policy Learning: Efficient Experimentation with Rich Outcomes

Zhiheng Zhang

Shanghai University of Finance and Economics

This talk presents a unified view of subsampling for causal learning, moving from estimating treatment effects with limited experimental samples to learning policies under rich, distribution-valued outcomes. Unlike standard subsampling, where a query only reveals a label, causal subsampling often requires an intervention, so the central question becomes where scarce experimental evidence should be collected. I will first discuss finite-sample active sampling for treatment effect estimation, then show how biased observational logs can guide randomized experimentation without compromising causal identification. Finally, I will connect these ideas to offline policy learning where each potential outcome is not a scalar but a probability distribution, and the policy is evaluated through a utility of the induced Wasserstein barycenter. This perspective shows that subsampling can support not only efficient estimation, but also reliable policy optimization when decisions affect the entire outcome distribution rather than just its mean.

Axis-Neighbor Geometry and Length-Scale Identifiability in Gaussian Process Models

◆ *Chenlu Shi, Lin Wang*

New Jersey Institute of Technology, Purdue University

In computer experiments, experimental designs for Gaussian process (GP) surrogates are often chosen to achieve global space-filling coverage for prediction. However, the same simulator runs must also support estimation of covariance hyperparameters, and designs that are effective for prediction alone may yield weak or unstable length-scale inference. This issue is especially acute for anisotropic GP models, where each coordinate has its own length-scale and informative comparisons are inherently directional. We develop a geometry-driven framework that links experimental design directly to length-scale identifiability in anisotropic GP models with Gaussian correlation. For general designs, we derive lower bounds for the diagonal Fisher information that depend explicitly on local axis-neighbor geometry through the minimum coordinate-isolating spacing and the multiplicity of closest such pairs. These results show that, in the small-length-scale regime, stable inference is governed by local directional structure rather than by global spread alone. Motivated by this theory, we propose a structured design that is a strong orthogonal array of strength 2+, and hence combines space-filling properties in one- and two-dimensional projections with a rich collection of short coordinate-isolating pairs. We further derive explicit Fisher-information bounds and sharp small-length-scale asymptotics for the proposed design, and introduce a structured level-expansion that refines global resolution while preserving the local geometry responsible for length-scale identifiability. Numerical studies indicate that the proposed designs substantially improve the stability of length-scale estimation while maintaining competitive predictive performance across a range of length-scale regimes.

Weighted Confidence Set via Randomized Allocation

Lingyou Pang, Yunan Wu, ◆ *Chenglong Ye*

University of California Davis, Tsinghua University, University of Kentucky

We study how to aggregate multiple confidence sets while preserving nominal coverage and adapting to set quality. Existing aggregation methods often either require fixed, pre-specified weights or incur additional coverage loss when weights are estimated from data. We propose a randomized aggregation framework with data-adaptive weights that achieves exact $(1-\alpha)$ validity. Our theory establishes conditional and marginal coverage guarantees, clarifies when adaptive weighting is valid, and covers constant-weight and independent-weight special cases. The framework is broadly applicable to conformal and non-conformal confidence set construction and provides a practical route to adaptive aggregation without sacrificing coverage.

Non-Smooth Optimization in Distributed Learning Environments

Joseph Casey, Boxiang Wang, ◆ *Keren Li*

University of Alabama at Birmingham, The University of Iowa, University of Alabama at Birmingham

Non-smooth optimization arises in many statistical learning problems, for example, support vector machines. In distributed learning environments, such objectives present additional challenges because subgradient based procedures can be sensitive to data heterogeneity and often require frequent communication across nodes. These difficulties become more pronounced when data distributions or modeling structures vary substantially across participating sites.

This work studies non-smooth optimization in distributed machine learning through an extended representative learning (RepL) framework. Earlier RepL methods construct point representatives that summarize local datasets by matching specific functional properties. The present work introduces a second generation representative construction, referred to as clouded representatives. Instead of individual pseudo observations, clouded representatives form structured distributions that encode local statistical information conditioned on the current model state. These representative distributions provide a compact description of local data while preserving the structure of the underlying optimization problem.

The framework is illustrated using a canonical non-smooth learning problem: support vector machines. Clouded representatives approximate the empirical risk contributions of local datasets while enabling collaborative optimization across nodes under substantial heterogeneity in both data distributions and modeling objectives. The resulting approach provides a flexible mechanism for distributed learning with reduced communication burden and improved robustness to heterogeneity.

S153: Uncertainty Quantification in Modern Statistics and Machine Learning

Conformal Selective Prediction with General Risk Control

Tian Bai, ◆ *Ying Jin*

Stanford University, University of Pennsylvania

In deploying artificial intelligence (AI) models, selective prediction offers the option to abstain from making a prediction when uncertain about model quality. To fulfill its promise, it is crucial to enforce strict and precise error control over cases where the model is trusted. We propose Selective Conformal Risk control with E-values (SCoRE), a new framework for deriving such decisions for any trained model and any user-defined, bounded and continuously-valued risk. SCoRE offers two types of guarantees on the risk among ``positive'' cases in which the system opts to trust the model. Built upon conformal inference and hypothesis testing ideas, SCoRE first constructs a class of (generalized) e-values, which are non-negative random variables whose product with the unknown risk has expectation no greater than one. Such a property is ensured by data exchangeability without requiring any modeling assumptions. Passing these e-values on to hypothesis testing procedures, we yield the binary trust decisions with finite-sample error control. SCoRE avoids the need of uniform concentration, and can be readily extended to settings with distribution shifts. We evaluate the proposed methods with simulations and demonstrate their efficacy through applications to error management in drug discovery, health risk prediction, and large language models.

Seeing Through Correlations: Disentangled Feature Importance

◆ *Jin-Hong Du, Kathryn Roeder, Larry Wasserman*

The University of Hong Kong, Carnegie Mellon University, Carnegie Mellon University

Quantifying feature importance with valid statistical uncertainty is central to interpretable machine learning, yet classical model-agnostic methods often fail under feature correlation, producing unreliable attributions and compromising statistical inference. Existing approaches—such as Shapley values and leave-one-covariate-out—are designed for feature selection and vulnerable to correlation distortion, limiting their robustness on model interpretation. We introduce Disentangled Feature Importance (DFI), a model-agnostic framework that resolves these limitations by combining principled statistical inference with computational flexibility. DFI leverages entropic optimal transport to learn flexible disentanglement maps and provide an interpretable pathway for understanding how importance is attributed through the data's correlation structure. The framework generalizes to flow matching and differentiable loss functions, enabling statistically

valid importance assessment for black-box predictors in both regression and classification. We establish statistical inference theory, which enables valid confidence intervals and hypothesis testing with Type I error control. Empirical results on synthetic and biomedical datasets show that DFI delivers substantially higher statistical power than removal-based and conditional permutation methods, while maintaining robust, interpretable attributions under severe feature interdependence.

Parameter Inference via Nonlinear Partial Differential Equations Informed Gaussian Processes

◆ *Zhaohui Li, Shihao Yang, Jeff Wu*

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Georgia Institute of Technology, The Chinese University of Hong Kong, Shenzhen

Partial differential equations (PDEs) are widely used for description of physical and engineering phenomena. Some key parameters involved in PDEs, which represents certain physical properties with important scientific interpretations, are difficult or even impossible to be measured directly. Estimation of these parameters from noisy and sparse experimental data of related physical quantities is an important task. Many methods for PDE parameter inference involve a large number of evaluations of numerical solution of PDE through algorithms such as finite element method, which can be time-consuming especially for nonlinear PDEs. We propose a novel method for estimating unknown parameters in PDEs, called PDE-Informed Gaussian Process Inference (PIGPI). Through modeling the PDE solution as a Gaussian process (GP), we derive the manifold constraints induced by the (linear) PDE structure such that under the constraints, the GP satisfies the PDE. For nonlinear PDEs, we propose an augmentation method that transfers the nonlinear PDE into an equivalent PDE system linear in all derivatives that our PIGPI can handle. PIGPI can be applied to multi-dimensional PDE systems and PDE systems with unobserved components. The method completely bypasses the numerical solver for PDE, thus achieving drastic savings in computation time, especially for nonlinear PDEs. Moreover, the PIGPI method can give the uncertainty quantification for both the unknown parameters and the PDE solution.

A Conformal Test of Linear Models via Permutation-Augmented Regressions

Leying Guan

Yale School of Public Health

Permutation tests are widely recognized as robust alternatives to tests based on normal theory. Random permutation tests have been frequently employed to assess the significance of variables in linear models. Despite their widespread use, existing random permutation tests lack finite-sample and assumption-free guarantees for controlling type I error in partial correlation tests. To address this ongoing challenge, we have developed a conformal test through permutation-augmented regressions, which we refer to as PALMRT. PALMRT not only achieves power competitive with conventional methods but also provides reliable control of type I errors at no more than 2α, given any targeted level α, for arbitrary fixed designs and error distributions. We have confirmed this through extensive simulations.

Compared to the cyclic permutation test (CPT) and residual permutation test (RPT), which also offer theoretical guarantees, PALMRT does not compromise as much on power or set stringent requirements on the sample size, making it suitable for diverse biomedical applications. We further illustrate the differences in a long-Covid study where PALMRT validated key findings previously identified using the t-test after multiple corrections, while both CPT and RPT suffered from a drastic loss of power and failed to identify any discoveries.

S155: Matrix Data

High-Dimensional Spectral Enhancement Discriminant Analysis

Yonghan Zhang, Zhangni Pu, Lu Yan, ◆ *Jiang Hu*

Northeast Normal University, Northeast Normal University, Northeast Normal University, Northeast Normal University

Regularized linear discriminant analysis (RLDA) is a widely used tool for classification and dimensionality reduction, but its performance in high-dimensional settings can vary substantially with the underlying covariance structure. Existing theoretical analyses provide limited insight into how such structural features affect classification accuracy. To address this issue, we derive a non-asymptotic approximation to the RLDA misclassification rate and use it to characterize the structural effect of the covariance matrix on classification performance. Motivated by this analysis, we propose a new classifier, termed spectral enhancement discriminant analysis (SEDA), which improves classification by modifying the regularization strength along selected spiked directions. Building on a new eigenvector result from random matrix theory, we further derive an asymptotic characterization of the SEDA misclassification rate, together with a bias-correction method and a theoretically motivated parameter-selection rule. Simulation studies and real-data experiments demonstrate that SEDA can outperform existing LDA-based methods in both classification and dimensionality reduction tasks.

The Spurious Factor Dilemma: Robust Inference in Heavy-Tailed Elliptical Factor Models

Jiang Hu, Jiahui Xie, ◆ *Yangchun Zhang, Wang Zhou*

Northeast Normal University, National University of Singapore, Shanghai University, National University of Singapore

Standard methods for determining the number of factors often overestimate the true number when data exhibit heavy-tailed randomness, misinterpreting noise-induced outliers as genuine factors. This paper addresses this challenge within the framework of Elliptical Factor Models (EFM), which accommodate both heavy tails and potential non-linear dependencies common in real-world data. We demonstrate, both theoretically and empirically, that heavy-tailed noise generates spurious eigenvalues that mimic true factor signals. To distinguish these, we propose a novel methodology based on a fluctuation magnification algorithm. Under mild conditions, we show that, by magnifying perturbations, the eigenvalues associated with real factors exhibit significantly less fluctuation (stabilizing asymptotically) than spurious eigenvalues arising from heavy-tailed effects. We develop a formal testing procedure based on this principle and apply it to the problem of accurately selecting the number of common factors in heavy-tailed EFMs. Simulation studies and real data analysis confirm the effectiveness of our approach, particularly in scenarios with pronounced heavy-tailedness.

On High-Dimensional Adaptive Sign Test

◆ *Chi Yao, Guangming Pan*

Anhui University, Nanyang Technological University

Diagnostic checking for serial dependence remains a fundamental problem in multivariate time series analysis. In high-dimensional settings, classical Portmanteau tests typically suffer from power loss. Unweighted sign-based Portmanteau tests provide a robust alternative by projecting observations onto the unit sphere; however, their reliance on uniform weights across all lags dilutes signal strength under sparse serial dependence structures. We develop an adaptive sign-based Portmanteau test to address this inefficiency in high-dimensional models. By introducing a data-adaptive weighting scheme that assigns exponentially larger weights to lags with significant correlations, the proposed test effectively filters uninformative noise and maximizes sensitivity to sparse alternatives. Asymptotically, we establish a Central Limit Theorem (CLT) for the weighted statistic as $p, n \to \infty$. This theoretical framework relaxes standard constraints by accommodating general covariance matrices, permitting the maximum lag m to diverge, and allowing for flexible innovation distributions, including independent components and elliptical families. Finally, we provide a ratio-consistent estimator for the asymptotic variance, ensuring the test can be implemented directly without prior specification of the population covariance structure.

Consistent and Powerful CUSUM Change-Point Test for Panel Data with Changes in Variance

Wenzhi Yang

Anhui University

This talk investigates change-point of variance in panel data models with time series of α-mixing. Based on the cumulative sum (CUSUM) method and the individual differences, we construct a CUSUM test for panel data models to detect variance changes. Under the null hypothesis, we derive the limit distribution of this test, which can be used to detect the change-point of variance. Under the alternative hypothesis, the limit behavior of the CUSUM test is also derived. To validate the performance of the test, we conducted simulation analyses with Gaussian and Gamma errors. The results demonstrate that this testing method significantly outperforms existing approaches, particularly in detecting sparse variance changes. Finally, we conducted a practical case study using panel data from the Shanghai Shenzhen CSI 300 Index Components. Not only did we successfully identify the change-points of variance, but we also delved deeper into the underlying economic drivers behind these changes.

S032: Causal Learning in Modern Health and Survey Data

Mediation Analysis for Complex Survey Designs

Wei Hao

University of Michigan

Mediation analysis is an important statistical methodology for understanding the causal pathways linking exposures, mediators and outcomes. It has been widely applied across numerous research fields, including social sciences, environmental health sciences, epidemiology, and nutritional sciences and so on. Large-scale population studies frequently employ complex survey designs involving clustering, stratification, and sampling weights, such as the National Health and Nutrition Examination Survey (NHANES). However, most methodology developments in mediation analysis have focused on data obtained through simple random sampling, and limited work has addressed mediation analysis that appropriately accounts for complex survey designs.

Utilizing the R package “survey”, we propose a framework for estimating indirect and direct effects while incorporating survey design features. Statistical inference is conducted using bootstrap resampling methods. We performed a series of simulations studies to evaluate the performance of the proposed approach and compared it with conventional methods that ignore survey design components. Finally, we illustrate the proposed method through

an application to data from the NHANES study.

Debiased Machine Learning of Average Dose-Response Functions with General Instrumental Variables

Shuyuan Chen

Zhejiang University

Estimating causal effects of continuous treatments, such as dose-response functions, is common in practice. Classical methods typically assume fully observed confounders, an assumption often violated due to unmeasured confounding. In this talk, I will introduce a novel framework for local identification of dose-response functions using instrumental variables to mitigate such bias. Our approach introduces uniform regular weighting functions and covers the treatment space with a finite collection of open sets, enabling local identification within each region.

For estimation, we develop an augmented inverse probability weighting (AIPW) score under a debiased machine learning framework and provide practical guidance for adaptively specifying weighting functions from the data. We establish asymptotic properties when the dose-response function is estimated via kernel regression or empirical risk minimization.

Nonparametric Estimation of Subgroup Mediation Effects with Semi-Competing Risks Data

◆ *Zi Wang, Jialiang Li, Yu Cheng*

University of Electronic Science and Technology of China, National University of Singapore, University of Pittsburgh

A treatment may exert its influence on a nonterminal event (e.g., disease progression), which in turn can affect the occurrence of a terminal event (e.g., death). Alternatively, the treatment may directly impact the terminal event without acting through the nonterminal event. Our goal is therefore to evaluate both the mediational effect of the treatment operating through the nonterminal event and the direct effect of the treatment on the terminal event. However, the conventional definitions of natural direct effect and natural indirect effect are not appropriate for the semi-competing risks data structure, where time to a nonterminal event may be censored by a terminal event, but not vice versa. In this work, we adopt a principal stratification framework to define the natural direct and indirect effects within the "always diseased" stratum, and to define the total effect within both the "always diseased" and "never diseased" strata. We develop nonparametric estimators for these causal effects under suitable technical assumptions, establish their theoretical properties, and demonstrate good finite-sample performance through numerical studies. This paper provides a flexible approach for estimating natural causal mediation effects and offers valuable insights into mediation mechanisms in semi-competing risks settings.

A Copula-Based Model for Causal Effects with Interval-Censored Data under Semi-Competing Risks

◆ *Tao Sun, Xuezhu Qi*

Renmin University of China, Renmin University of China

Interval-censored data and semi-competing risks commonly arise in longitudinal survey studies, where the non-terminal event time is observed within intervals and the terminal event may preclude non-terminal outcomes. Their coexistence, along with complex dependence structures, poses significant challenges for causal inference. We propose a copula-based framework for causal inference with interval-censored semi-competing risks data. Under the potential outcomes framework, principal stratification is used to define causal effects on non-terminal and terminal events within meaningful subpopulations. A flexible copula model captures the dependence between event times and across potential outcomes. Under interpretable assumptions on censoring, treatment assignment, and dependence, we derive principal-stratum causal effects and develop estimation procedures. The proposed approach provides a flexible framework that accommodates both interval censoring and semi-competing risks, with broad applications in health and social research.

S011: Advances in High-Dimensional Dependence Modeling

Structural Identification for Spatial-Temporal Dynamic Models

Rongmao Zhang

Zhejiang Gongshang University

Identifying latent cluster structures in spatial trends constitutes an important yet challenging task in diverse applications. In this paper, we propose a novel method based on a discrepancy measure over small spatial blocks that effectively uncovers heterogeneity within dynamic spatial data. Our approach effectively detects boundaries where structural changes occur, thus allowing for more nuanced insights into underlying spatial patterns. Unlike methods predicated on strong stationarity assumptions, our framework accommodates piecewise-defined parameters and irregular sampling locations, enabling its broad applicability to real-world datasets. We further establish asymptotic properties and limit distributions of the proposed methods by leveraging the notion of spatial physical dependence, accounting for correlations across spatial domains. Simulations and real data analyses confirm the effectiveness of the method, highlighting its robustness and accuracy in identifying complex spatio-temporal structures.

Spatial Discrete Data Learning

Baisuo Jin

University of Science and Technology of China

Spatial discrete data learning plays a vital role in modern operations management, especially in studies where spatial relations significantly impact decision making. In this paper, we propose a generalized linear model with a link function that is not only dependent on the regressors but is also spatially autocorrelated to analyze non-normal spatial data. We develop a method to derive maximum likelihood type estimators of the parameters using Taylor expansion, thus avoiding solving nonlinear problems. The proposed method is conceptually simple and easy to implement. Under mild conditions, we show that the estimators obtained by two-stage maximum likelihood type estimation are consistent and asymptotically normally distributed. In addition, we propose a weight matrix selection method and a model averaging approach to address uncertainty in spatial structure. Extensive simulation studies and a real data example demonstrate the effectiveness of the proposed method.

Estimation and Inference for High-Dimensional Spatial Asset Pricing Models with Heterogeneous Interactions

◆ *Zhaoxing Gao, Sihan Tu, Ruey Tsay*

University of Electronic Science and Technology of China, Zhejiang University, University of Chicago

This paper studies estimation and inference for a class of high-dimensional spatial factor models with unit-specific spatial interactions. The framework accommodates both observable and latent factors and is designed for panel systems that exhibit structured cross-sectional dependence together with common factor-driven comovement. We consider a high-dimensional regime in which both the cross-sectional dimension and the sample size may diverge. For models with observable factors, we propose a shrinkage Yule-Walker estimator that combines low-order factor moment conditions with ridge regularization. The method is computationally attractive when likelihood-based procedures may become burdensome or unstable, particularly in the presence of heterogeneous spatial coefficients. For models with latent factors, we develop a two-step procedure that first recovers the dynamic factor space through autocovariance-based eigenanalysis and then estimates the structural parameters using the same ridge-regularized Yule-Walker approach. We establish asymptotic results for the estimated spatial coefficients, loading matrices, and latent factors under suitable regularity conditions. Simulation experiments show that the proposed method performs well in finite samples. Empirical applications to U.S. stock returns and housing prices further {\color{red} show that the proposed} method delivers competitive, and often lower, out-of-sample forecast errors relative to benchmark alternatives.

Covariate Shift Correction for Spatial Autoregressive Model: A Reweighting Approach

◆ *Li Hou, Baisuo Jin, Ching-Kang Ing*

University of Science and Technology of China, University of Science and Technology of China, Tsing Hua University, Taiwan

This paper investigates parameter estimation in Spatial Autoregressive (SAR) models under covariate shift, where the distributions of training and test covariates differ. Such distributional shifts, often compounded by model misspecification, can lead to inconsistent estimates and reduced predictive accuracy in standard spatial econometric methods.

To address this, we propose a novel least-squares estimation method based on the eigendecomposition of the spatial weights matrix (EDLS), integrated with an importance-weighting scheme. By leveraging eigendecomposition for dimensionality reduction and reweighting the training data using trimmed density ratios, our approach effectively mitigates domain discrepancies. The importance-weighted EDLS estimator achieves rapid convergence in terms of prediction error and preserves consistency under model misspecification, ultimately converging to the pseudo-true parameter. Numerical results further demonstrate that our method enhances the model's generalization ability and improves out-of-sample prediction performance.

S018: Advances in Statistical Machine Learning Methods for Complex Data

Functional Auxiliary Information Synthesis in Survival Analysis

◆ *Jie Ding, Jialiang Li, Ingrid Van Keilegom, Xiaoguang Wang*

Dalian Maritime University, National University of Singapore, Katholieke Universiteit Leuven, Dalian University of Technology

Fusing related auxiliary information is promising in enhancing statistical performance of target tasks. We address the challenge of fitting survival models with scarce target time-to-event data in the presence of auxiliary information that is function-valued, as opposed to Euclidean statistics or individual-level records from sources that dominate existing literature. A novel methodology that transfers model-free estimators for survival functions across multiple subgroups is developed to improve the assessment of prognostic factors of interest. The bridge between target and auxiliary information is constructed through asymptotically vanishing functions, which are further linked to the classical target-only estimator via data-driven sieve-approximated leveraging functions. Our asymptotic analysis clarifies when and to what extent auxiliary information improves target estimation, and demonstrates explicitly that our estimator achieves asymptotic efficiency gains over conventional methods. Extensive simulations support these findings and a real-world case study

highlights the practical feasibility of our approach.

Fair Regression in Reproducing Kernel Hilbert Spaces: Single-Machine and Decentralized Implementations under Conditional Mean Parity

Xiaojun Mao

Shanghai Jiao Tong University

Machine learning has been widely applied across diverse domains, offering powerful tools for data processing and decision-making. Yet its use also raises serious concerns about fairness, as automated systems may introduce discriminatory biases and critical ethical challenges. This paper investigates the fair regression problem under the notion of Conditional Mean Parity (CMP) framework, which requires statistically balanced outputs across sensitive subgroups when conditioned on target variable intervals. We extend CMP to Reproducing Kernel Hilbert Spaces, formulating fair kernel (ridge) regression (FKR/FKRR) problems that jointly optimize performance and fairness. To improve scalability, we further propose a decentralized variant, decentralized fair kernel ridge regression (DFKRR) using random feature approximations, and design an efficient DFKRR-ADMM algorithm for distributed optimization. Our theoretical analysis establishes key properties of CMP and clarifies its connections with existing fairness notions. We also derive fairness performance trade-off bounds for both FKR and FKRR, providing theoretical guarantees for their balanced behavior. Moreover, we prove that the proposed algorithm achieves linear convergence, with DFKRR attaining generalization error bounds identical to those of its single-machine counterpart FKRR. Comprehensive experiments on synthetic and real-world datasets corroborate our theory, demonstrating superior fairness performance trade-offs compared with state-of-the-art baselines.

Integral Probability Metric-Guided CUSUM-Net for Nonparametric Changepoint Detection

Guanghui Wang

Nankai University

We propose CUSUM-Net, a nonparametric method for changepoint detection based on integral probability metrics (IPMs) and deep neural networks. Our approach learns a critic function by maximizing an aggregate CUSUM objective over candidate changepoints, thereby linking changepoint detection to two-sample IPM optimization. The learned critic induces a one-dimensional representation on which changepoints are localized by a classical CUSUM scan. Unlike parametric procedures, CUSUM-Net accommodates complex, high-dimensional distributional changes and applies to a range of data modalities, including Euclidean data, symmetric positive-definite matrices, images, and graphs. We establish excess-risk bounds for the learned critic under Hölder smoothness assumptions, with faster rates when the data exhibit low-dimensional manifold structure, and we derive corresponding changepoint localization guarantees. Numerical experiments demonstrate the flexibility and effectiveness of the proposed method.

Variational Autoencoding Discrete Diffusion with Enhanced Dimensional Correlations Modeling

◆ *Tianyu Xie, Shuchen Xue, Zijin Feng, Tianyang Hu, Jiacheng Sun, Zhenguo Li, Cheng Zhang*

Peking University, Chinese Academy of Sciences, Huawei Foundation Model Department, The Chinese University of Hong Kong, Shenzhen, Huawei Foundation Model Department, Huawei Foundation Model Department, Peking University

Discrete diffusion models have recently shown great promise for modeling complex discrete data, with masked diffusion models (MDMs) offering a compelling trade-off between quality and generation speed. MDMs denoise by progressively unmasking multiple dimensions from an all-masked input, but their performance can degrade when using few denoising steps due to limited modeling of inter-dimensional dependencies. In this paper, we propose Variational Autoencoding Discrete Diffusion (VADD), a novel framework that enhances discrete diffusion with latent variable modeling to implicitly capture correlations among dimensions. By introducing an auxiliary recognition model, VADD enables stable training via variational lower bounds maximization and amortized inference over the training set. Our approach retains the efficiency of traditional MDMs while significantly improving sample quality, especially when the number of denoising steps is small. Empirical results on 2D toy data, pixel-level image generation, and text generation demonstrate that VADD consistently outperforms MDM baselines in sample quality with few denoising steps.

S037: Decoding Complexity: Genomic Diversity, Disease Mechanisms, and Evolutionary Insights

Radiogenomics of Smoking-Related Lung Disease: CT Radiomics Integrates with Airway Transcriptomics

Fenghai Duan

Brown University

Objective: To identify biological distinctions across smoking-related lung disease phenotypes by integrating CT radiomics with airway (bronchial epithelial) transcriptomics alongside clinical data.

Methods: We applied k-means clustering to quantitative

CT radiomic features to define imaging-based subgroups, then compared clinical phenotypes across clusters. Airway transcriptomic profiles from RNA-seq were interrogated using differential expression, gene-set enrichment, and expression-variation analyses.

Results: Three radiomic clusters emerged—preserved, interstitial-predominant, and emphysema-predominant. Relative to the preserved cluster, the interstitial and emphysema clusters exhibited poorer lung function, lower exercise capacity, and worse quality of life, and experienced faster functional decline, greater emphysema progression, more respiratory events, and higher mortality. Outcomes were most severe in the emphysema-predominant cluster, followed by the interstitial-predominant cluster. Airway transcriptomic analyses linked severe radiomic phenotypes to heightened inflammatory signaling—particularly TNF-α–related pathways—whereas milder disease showed up-regulation of T-cell–associated genes.

Conclusions: Radiogenomic integration of CT radiomics and airway transcriptomics identifies three clinically and biologically distinct subgroups among heavy smokers. The findings connect clinical severity to airway inflammation, implicating TNF-α–mediated pathways as potential targets for risk stratification and intervention.

Predicting Influenza Virus Evolution with Site-Based Mutation Dynamic Model

◆ *Maggie Haitian Wang, Chris Mok, Lirong Cao*

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

Prediction of virus evolution is crucial for pandemic and epidemic preparedness. In this talk, we will introduce a novel framework, beth-1, to forecast the genetic evolution of the influenza virus. Unlike traditional phylogenetic methods, this method models evolution from a site-based perspective, learning the fitness dynamics of mutations from global genomic and epidemiological data. Through retrospective, prospective, and experimental validations, we demonstrated that the proposed method offers more accurate predictions of future virus populations compared to existing approaches.

Estimation of Direct and Indirect Polygenic Effects and Gene-Environment Interactions Using Polygenic Scores in Case-Parent Trio Studies

◆ *Ziqiao Wang, Nilanjan Chatterjee*

University of Virginia, Johns Hopkins University

We propose PGS-TRI, a framework for analyzing polygenic scores (PGS) in case-parent trio studies that estimates the risk of an index condition associated with direct PGS effects, gene-environment interactions, and asymmetric maternal and paternal indirect effects. Simulations confirm its robustness in the presence of complex population structure and assortative mating. Applied to multi-ancestry autism spectrum disorders (ASD) trios (Ntrio = 18,383), PGS-TRI yielded transmission-based direct effects of PGSs for ASD and other neurocognitive traits along a genetic ancestry continuum, and identified asymmetric indirect effects of parental PGS for BMI and neurocognitive traits on children's ASD risk. In a trio study of European and Asian orofacial clefts (OFCs) (Ntrio = 1,904), PGS-TRI estimated direct and indirect effects of an established PGS and its interaction with maternal risk factors. Finally, we applied PGS-TRI to large-scale transcriptome-wide and metabolome-wide traits to examine their direct and indirect effects on ASD and OFCs risk.

Refining Structural Variation-Trait Associations with SVrefiner

◆ *Xiaoxuan Xia, Junteng Wu, Zhezhi Gao, Peixiong Yuan, Sheng Ye*

Tianjin University, Tianjin University, Tianjin University, Tianjin University, Tianjin University

Structural variants (SVs) are a major source of genetic diversity and disease risk across species. Despite advances in SV discovery enabled by graph-based pangenomes, most single-nucleotide polymorphism (SNP)–centric frameworks overlook SV sequence detail, treat overlapping SVs as independent variants, and consequently lose power in genome-wide association studies, bias heritability estimation, and degrade risk-prediction accuracy. Here we present SVrefiner, a scalable, sequence-congruence–guided algorithm that resolves overlapping SVs into a canonical set of non-overlapping refined SVs (rSVs), yielding high-fidelity genotype matrices for downstream analyses. Applied across human, plant, and animal genomes, SVrefiner uncovered tens of thousands of rSVs and revealed that discrete SV subregions, rather than entire events, frequently drive gene-expression traits. Incorporating rSVs increased the proportion of heritability explained, improved risk-prediction accuracy, and boosted cis-eQTL discovery. Benchmark evaluations further showed substantial computational efficiency, with the greatest advantages for longer SVs. By integrating sequence context into SV analysis, SVrefiner transforms how structural variants are represented, interpreted, and leveraged, establishing a new standard for trait and disease association studies in the pangenome era.

S043: From Data to Biomarkers: Statistical Innovation for Complex Challenges in Modern Genetics, Genomics and Proteomics

Horizontal Pleiotropy in Mendelian Randomization: Use It If You Cannot Beat It!

Anagh Chattopadhyay, ◆ *Nilanjan Chatterjee*

Johns Hopkins University

Mendelian randomization (MR) is a widely used approach for estimating causal effects of exposures on outcomes using genetic variants as instrumental variables; however, existing methods remain vulnerable to bias and/or loss of power in the presence of invalid instruments. We hypothesize that closely related outcome traits are likely to have a large overlap in underlying valid instruments in relation to a given exposure and that instrument borrowing (IB) across such traits can therefore yield more robust MR inference. To operationalize this idea, we introduce a novel co-heterogeneity statistic, computable from SNP-level ratio estimates for a given set of instruments associated with an exposure, to identify secondary outcome traits that are likely to share common root(s) for instrument invalidity with the primary trait. We then propose two new methods—IB-Mode and IB-PRESSO—which extend the popular MR approaches MRMode and MR-PRESSO to leverage the secondary trait(s) to improve MR analysis for the primary trait. Extensive simulation studies demonstrate that IB-based methods consistently outperform their standard counterparts, with gains increasing as the degree of overlap in valid instruments across outcome traits grows. Applications to established positive and negative control hypotheses suggest that IB-based methods offer improved control of type I error and increased power, with efficiency gains sometimes observed in the range of 100–200\%. We further illustrate the practical utility of these methods by revisiting the long-debated hypothesis on the causal effect of vitamin D on cardiometabolic traits.

Statistical Tools for Integrative Proteogenomic Analysis

Pei Wang

Icahn School of Medicine at Mount Sinai

In contemporary proteogenomic cancer studies, a single tumor is often analyzed with more than ten different proteogenomic profiling strategies, ranging from genetics, transcriptomics, proteomics, and post-translational modifications. In this talk, I will discuss the principal needs and challenges of multi-omics integration, with a focus on characterizing the functional consequences of genomic alterations through systematic cis-, trans-regulatory, and network analyses. I will illustrate how these approaches illuminate the regulatory effects of mutations, copy number changes, and methylations. Analyses of recent CPTAC proteogenomic datasets illustrate practical workflows, tool usage, and performance considerations in real-world data contexts.

Quantile-Based Discriminant Analysis for High-Dimensional Imbalanced Classification

Qian Tang, Yuwen Gu, ◆ *Boxiang Wang*

University of Minnesota, University of Connecticut, University of Iowa

I will present the quantile-based discriminant analysis (QuanDA), a new method for addressing class imbalance in high-dimensional classification problems. Inspired by the least squares view of linear discriminant analysis, QuanDA employs the quantile regression, enhanced with elastic-net regularization, for feature selection and classification. Unlike traditional approaches to discriminant analysis, such as the linear, quadratic, and Fisher discriminant analyses, QuanDA does not rely on heavy model assumptions, thus significantly increasing the robustness of the classification. I will present comprehensive simulation studies and benchmark tests to demonstrate that QuanDA achieves competitive performance relative to other leading classification methods that deal with class imbalance.

Using Multiple Biomarkers for Patient Enrichment in Two-Stage Clinical Designs

Fengqing Zhang

Drexel University

Enrichment strategies play a critical role in modern clinical trial design, especially as precision medicine advances the focus on patient-specific efficacy. By targeting biomarker-defined populations most likely to benefit, these strategies improve efficiency, reduce sample sizes and costs, accelerate timelines, and minimize unnecessary exposure to treatment-related risks and side effects. Recent developments in enrichment design have introduced biomarker randomness and accounted for the correlation structure between treatment effect and biomarker, resulting in a two-stage threshold enrichment design. However, existing methods are limited to a single continuous biomarker. While incorporating multiple biomarkers can improve the accuracy of target population identification, no study has examined how to incorporate multiple continuous biomarkers into enrichment designs due to the challenges in determining multiple thresholds. To fully utilize information from all relevant biomarkers, we propose novel two-stage enrichment designs capable of handling two or more continuous biomarkers. Our framework accommodates two popular treatment effect metrics including average treatment effect (ATE) and the standardized ATE. We illustrate our method using a hypothetical clinical trial involving early-stage Alzheimer's patients and assess the impact of stage one sample size on threshold estimation through a simulation study. Overall, our proposed two-stage enrichment designs offer researchers greater flexibility in integrating multiple continuous biomarkers. The findings from our study provide valuable insights for the advancement of enrichment trial methodology.

S044: From Probabilistic Theory to Causal Inference in Complex Systems

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

Alexandre Belloni, Ethan Fang, ◆ *Shuting Shen*

Duke University, Duke University, National University of Singapore

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.

Estimating Treatment Effect Sizes Without Knowing Who Was Treated

Yao Zhang

National University of Singapore

We develop model-assisted randomization tests for time-series experiments with panel data. To avoid sample splitting when building a treatment effect model, we propose a new estimation method that learns the sizes of lagged conditional average treatment effects (CATEs) from the covariance structure of residualized outcomes. Assuming the sign of the CATEs is known, we then use these estimates to construct powerful test statistics and conditional randomization p-values for

testing sharp and lag-specific null hypotheses. We show that the resulting CATE estimators improve the power of randomization tests compared with tests that use no model or rely solely on outcome models.

Mediation Analysis with Interval-Censored Outcome and Interval-Censored Mediator

◆ *Dongdong Li, Rui Wang*

Harvard University, Harvard University

Existing methods for mediation with survival outcomes assume the mediator is fully observed. In many settings, however, both the mediator and the outcome are interval-censored because they are detected only at periodic assessments. We propose a mediation analysis framework for doubly interval-censored data by leveraging a joint proportional hazards model estimated via an EM algorithm. From the fitted model we compute counterfactual survival curves, natural indirect and direct effects, the proportion mediated as a function of time, as well as restricted mean survival time decompositions. We apply the method to the Zurich Primary HIV Infection cohort to evaluate whether time to viral suppression mediates the effect of sex on time to viral rebound after antiretroviral therapy interruption.

S051: High Dimensional Data Analysis and Change Point Inference

Two-way Node Popularity Model for Directed and Bipartite Networks

Jiangzhou Wang

Shenzhen University

There has been extensive research on community detection in directed and bipartite networks. However, existing studies often overlook node popularity across different communities, a common phenomenon in real-world networks. To address this issue, we propose a new probabilistic framework, called the Two-Way Node Popularity Model (TNPM), which accommodates edges from a general sub-Gaussian family. We develop the Delete-One-Method (DOM) for model fitting and community structure identification, and provide a comprehensive theoretical analysis with new techniques for sub-Gaussian generalization. We also propose the Two-Stage Divided Cosine Algorithm (TSDC) for large-scale networks. Extensive numerical studies demonstrate the advantages of the proposed methods in estimation accuracy and computational efficiency. Two real-world applications are further presented to illustrate the usefulness of our methods.

Adaptive Change Point Inference for High Dimensional Time Series with Temporal Dependence

◆ *Xiaoyi Wang, Jixuan Liu, Long Feng*

Beijing Normal University, Nankai University, Nankai University

This paper investigates change point inference in high-dimensional time series. We begin by introducing a max-L_2-norm based test procedure, which demonstrates strong performance under dense alternatives. We then establish the asymptotic independence between our proposed statistic and the two max-L_∞-based statistics introduced by Wang (2023). Building on this result, we develop an adaptive inference approach by applying the Cauchy combination method to integrate these tests. This combined procedure exhibits robust performance across varying levels of sparsity. Extensive simulation studies and

real data analysis further confirm the superior effectiveness of our proposed methods in the high-dimensional setting.

Simultaneous Detection and Localization of Mean and Covariance Changes in High Dimensions

◆ *Junfeng Cui, Guangming Pan, Guanghui Wang, Changliang Zou*

Shenzhen University, Nanyang Technological University, Nankai University, Nankai University

Existing methods for high-dimensional changepoint detection and localization typically focus on changes in either the mean vector or the covariance matrix separately. This separation reduces detection power and localization accuracy when both parameters change simultaneously. We propose a simple yet powerful method that jointly monitors shifts in both the mean and covariance structures. Under mild conditions, the test statistics for detecting these shifts jointly converge in distribution to a bivariate standard normal distribution, revealing their asymptotic independence. This independence enables the combination of the individual p-values using Fisher's method, and the development of an adaptive p-value-based estimator for the changepoint. Theoretical analysis and extensive simulations demonstrate the superior performance of our method in terms of both detection power and localization accuracy.

Adaptive L-statistics for high dimensional test problem

◆ *Huifang Ma, Long Feng, Zhaojun Wang*

Nankai University, Nankai University, Nankai University

In this study, we focus on applying L-statistics to the high-dimensional one-sample location test problem. Intuitively, an L-statistic with k parameters tends to perform optimally when the sparsity level of the alternative hypothesis matches k. We begin by deriving the limiting distributions for both L-statistics with fixed parameters and those with diverging parameters. To ensure robustness across varying sparsity levels of alternative hypotheses, we first establish the asymptotic independence between L-statistics with fixed and diverging parameters. Building on this, we propose a Cauchy combination test that integrates L-statistics with different parameters. Both simulation results and real data applications highlight the advantages of our proposed methods.

S064: Model Averaging and Related Topics

Mallows-Type Model Averaging: Non-Asymptotic Analysis and All-Subset Combination

Jingfu Peng

Tsinghua University

Model averaging (MA) and ensembling play a crucial role in statistical and machine learning practice. When multiple candidate models are considered, MA techniques can be used to weight and combine them, often resulting in improved predictive accuracy and better estimation stability compared to model selection (MS) methods. In this paper, we address two challenges in combining least squares estimators from both theoretical and practical perspectives. We first establish several oracle inequalities for least squares MA via minimizing a Mallows' C_p criterion under an arbitrary candidate model set. Compared to existing studies, these oracle inequalities yield faster excess risk and directly imply the asymptotic optimality of the resulting MA estimators under milder conditions. Moreover, we consider candidate model construction and investigate the problem of optimal all-subset combination for least squares estimators, which is an important yet rarely discussed topic in the existing literature. We show that there exists a fundamental limit to achieving the optimal all-subset MA risk. To attain this limit, we propose a novel Mallows-type MA procedure based on a dimension-adaptive C_p criterion. The implicit ensembling effects of several MS procedures are also revealed and discussed. We conduct several numerical experiments to support our theoretical findings and demonstrate the effectiveness of the proposed Mallows-type MA estimator.

A scalable model averaging based on Kullback-Leibler distance for multivariate regression models

Jie Zeng

Hefei Normal University

This paper considers estimation problem in multivariate regression models. Under this framework, we develop a novel two-stage model averaging procedure. In the first stage, we construct a scalable model averaging estimator which involves transforming the original model based on the singular value decomposition. When the dimension of the regressor vector is K, this approach enables us to average the estimators from the candidate model set of size K instead of size 2^{K}. The second stage is to find the optimal weights for averaging by applying a weight choice criterion from Kullback-Leibler distance. We prove that the minimum loss from the scalable model averaging is asymptotically the same as that from original model averaging, further demonstrate asymptotic optimality of the scalable model averaging estimator using Kullback-Leibler-distance-based weights, and derive the rate of the resulting weights tending to the risk-based optimal weights. In comparison with existing model averaging methods, the simulation results show that, in terms of mean squared prediction error and computation time, our proposal is more efficient, especially under the situation where

the number of candidate models is large and the sample size is small. Moreover, a real data analysis is provided to illustrate the application of our method in practice.

Handling Covariate Shift by Model Averaging

◆ *Yifan Zhang, Xinyu Zhang*

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

This paper develops and analyzes an adaptive importance-weighted model averaging (AIWMA) method for prediction under covariate shift. The goal is to construct a prediction rule with low prediction risk under the target distribution, using labeled source data and unlabeled target covariates when the source and target covariate marginal distributions may differ. The unweighted empirical risk minimizer is based on the source covariate distribution and may therefore perform poorly under the target distribution. Importance weighting by the density ratio between the target and source covariate distributions provides a natural correction. However, the resulting importance-weighted estimator can be unstable in finite samples because the weighted empirical criterion may be dominated by a small number of observations with large density ratios. These considerations motivate a model-averaging approach in which the strength of importance-weighting correction is treated as a source of model uncertainty rather than fixed in advance. We therefore average a family of importance-weighted least-squares estimators indexed by different correction levels. The averaging weights are selected by a data-driven criterion designed to estimate the prediction risk under the target distribution. Asymptotic optimality and weight concentration are proved for the proposed method under regularity conditions. Simulation studies and a Parkinson's disease telemonitoring analysis show that the proposed averaging procedure achieves favorable predictive performance under the target distribution and often improves upon standard importance weighting and model-selection-based alternatives.

Trans-MA: Sufficiency-principled Transfer Learning via Model Averaging

Xiyuan Zhang, ◆ *Huihang Liu, Xinyu Zhang*

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Shanghai University of Finance and Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Domain aggregation in multi-source transfer learning faces a critical challenge: effectively integrating knowledge from heterogeneous sources while addressing statistical uncertainties. Existing methods rely on restrictive single-similarity assumptions (i.e., individual or combinatorial similarity) and often neglect practical variability, leading to suboptimal performance. To address these limitations, we propose a sufficiency-principled transfer learning framework that systematically balances model averaging and model selection during domain aggregation with unknown informative knowledge. The framework employs a sufficiency principle for quantifying transferable knowledge to eliminate the challenges of spurious correlation and perturbated evaluation. The proposed model averaging algorithms accommodate both individual and combinatorial similarity regimes, and also has privacy-preserving mechanisms. Theoretically, we establish the asymptotic optimality, estimator convergence and asymptotic normality, for multiple source domain linear regression models with diverging parameters. Especially, compared with existing results, we provide enhanced rate of converge for parameter of interest. Empirical validation through extensive simulations and an analysis of Beijing housing rental data demonstrates the statistical superiority of our framework over conventional domain aggregation methods. The proposed methodology extends beyond regression models, offering a generalizable paradigm for transfer learning in statistical decision theory.

S065: Model Averaging for Time Series Data

Model Averaging for Functional Quantile Regression

◆ *Wenchao Xu, Xinyu Zhang, Jeng-Min Chiou, Yuying Sun*

Shanghai University of International Business and Economics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Taiwan University & Academia Sinica, Taiwan, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

In data-rich environments, functional data from various domains are employed to forecast conditional quantiles. However, the infinite-dimensional nature of functional data introduces uncertainty. To address this uncertainty, this paper proposes a novel conditional quantile model averaging approach that selects the data-driven weights by minimizing a K-fold cross-validation. Two novel theoretical justifications are provided. First, when all candidate models are misspecified, the selected weights are shown to be asymptotically optimal for minimizing excess final prediction error which removes the dominant innovation term, and the convergence rate is derived. Second, when the correctly specified model belongs to the candidate set, a convergence rate for the averaged estimators is established. Simulation studies and empirical applications to forecasting extreme quantiles in cryptocurrency markets demonstrate that the proposed approach frequently outperforms other model selection and averaging techniques.

Model Uncertainty and Averaging of Spatial Autoregression

for Spatial Heterogeneity

Yuge Luo, Wangli Xu, Zudi Lu, ◆ Jun Liao

Renmin University of China, Renmin University of China, City University of Hong Kong, Renmin University of China

The prevalence of spatial heterogeneity and dependence introduces substantial challenges for non-linear spatial data modeling,particularly because of the presence of model uncertainty. To address these issues,we develop a novel model averaging framework within a semiparametric family of spatial autoregressive models with partially varying coefficients (SAR-PVC).In this article,we propose a blockwise cross-validation (BCV) criterion for selecting model averaging weights,thereby mitigating uncertainty arising from ambiguity over model specifications such as which coefficients vary. Incorporating structural information of the coefficients,we first devise a profile-likelihood-based estimation method for SAR-PVC parameters,and establish the asymptotic distribution of the resulting estimators under correct model specification. Moreover,even under misspecified models,we prove the consistency of the weight estimators,and demonstrate the asymptotic optimality of the resulting model averaging estimator. Monte Carlo simulations and illustrative examples further confirm the effectiveness of the proposed SAR-PVC estimators and the BCV-based model averaging procedure.

Data Decomposition-Based Model Averaging

Yongmiao Hong, ◆ Wencan Lin, Yuying Sun, Shouyang Wang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

We develop a general econometric framework for forecasting macroeconomic and financial time series that integrates time series decomposition, machine learning, and model averaging. By exploiting frequency heterogeneity in predictors, the framework constructs machine learning models for distinct frequency components of predictors to capture heterogeneous market dynamics. Theoretically, the asymptotic optimality of the forward-validation model averaging estimator is established; the estimator is shown to attain the minimum out-ofsample prediction risk in large samples, even when all candidate models are misspecified. Empirically, the framework is applied to forecast major exchange rates using a broad set of macroeconomic and financial predictors, yielding statistically significant out-of-sample predictive gains relative to the random walk benchmark. The weight dynamics also reveal an interpretability mechanism: short-horizon predictability emerges from balanced high- and low-frequency contributions with a slight advantage from high-frequency variation, while long-horizon predictability is dominated by low-frequency macroeconomic fundamentals.

Optimal Parameter-Transfer Learning by Time-Varying Model Averaging

Boyan Li, ◆ Yuying Sun, Xinyu Zhang

Chinese Academy of Sciences, Chinese Academy of Sciences, Chinese Academy of Sciences

The presence of structural changes in economics and related fields poses challenges for outof-sample forecasting, particularly when the sample size of the target data is limited. This paper proposes a novel optimal parameter transfer learning approach through time-varying model averaging (Transfer-TVMA), which enhances target model predictions by adaptively transferring potentially shared parameter information under structural changes and model uncertainty. A local forward-validation weight choice criterion is developed to select time-varying combination weights for candidate models based on different populations. The asymptotic optimality and convergence properties of the selected time-varying weights are established under regularity conditions. Beyond point forecasts, we further develop a conformal prediction interval algorithm based on TransferTVMA and provide its asymptotic validity without the exchangeability assumption. Simulation studies and empirical applications to inflaction forecasting and exchange rate forecasting demonstrate the superior predictive performance of the proposed approach compared to existing competing methods.

S072: Modern Statistical Methods and Artificial Intelligence: From Theory to Biomedical Applications

The Future of AI Scientists: Emerging Directions and Fundamental Challenges

Jie Ding

University of Minnesota

Modern AI systems are rapidly evolving from passive tools into agentic systems that can reason, learn, collaborate, and interact with human scientists in the research process. This talk discusses emerging directions in AI scientist systems, drawing inspiration from human-centered principles such as lifelong learning, teaching, collaboration, objective alignment, and auditing. The goal is to identify fundamental research challenges and stimulate new perspectives on building AI systems that can operate reliably, adaptively, and transparently in complex real-world research settings.

Mismatch-Aware Transfer Learning for Optimal Linear Pre-

diction

Jinhong Li, ◆ *Lu Tang, Guan Yu*

University of Pittsburgh, University of Pittsburgh, University of Pittsburgh

Transfer learning can improve prediction performance in a target study by leveraging data from related external source studies. Its effectiveness is often limited by two challenges: (i) distributional heterogeneity across studies and (ii) covariate mismatch where only subsets of predictors from the target study are available in the source studies. In this paper, we first tackle posterior drift by introducing MATOL (Mismatch-Aware Transfer Learning for Optimal Linear Prediction), a second-order moment-based estimator that uses covariate covariance and covariate-outcome cross-covariance to obtain the optimal linear predictor for the target population. MATOL selectively borrows information from external studies by adaptively fusing heterogeneous covariate-outcome cross-covariance without imputing unobserved covariates. Next, we develop extensions of MATOL that can further (i) accommodate external studies with heterogeneous covariate distributions, and (ii) maintain robustness to heavytailed noise via Huber's M-estimation. Unlike standard transfer-learning pipelines and conventional pooling approaches that typically assume fully aligned and observed covariates, MATOL is explicitly designed to handle mismatched covariate settings. Extensive simulation studies demonstrate the effectiveness of MATOL and strong robustness under heterogeneity. We further apply MATOL to two real-world prediction tasks, ICU length of stay for sepsis patients and treatment continuation using insurance claims, where it consistently outperforms both imputation-based and pooled baselines in out-of-sample prediction.

Hierarchical Contrastive Learning for Multimodal Data with Partial Sharing

Huichao Li, Junhan Yu, ◆ *Doudou Zhou*

National University of Singapore

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.

High-Dimensional Covariate-Augmented Overdispersed Poisson Factor Model with Application to Single-Cell Multi-Omics

◆ *Wei Liu, Qingzhi Zhong*

Sichuan University, Jinan University

The current Poisson factor models often assume that the factors are unknown, which overlooks the explanatory potential of certain observable covariates. This study focuses on high dimensional settings, where the number of the count response variables and/or covariates can diverge as the sample size increases. A covariate-augmented overdispersed Poisson factor model is proposed to jointly perform a high-dimensional Poisson factor analysis and estimate a large coefficient matrix for overdispersed count data. A group of identifiability conditions are provided to theoretically guarantee computational identifiability. We incorporate the interdependence of both response variables and covariates by imposing a low-rank constraint on the large coefficient matrix. To address the computation challenges posed by nonlinearity, two high-dimensional latent matrices, and the low-rank constraint, we propose a novel variational estimation scheme that combines Laplace and Taylor approximations. We also develop a criterion based on a singular value ratio to determine the number of factors and the rank of the coefficient matrix. Comprehensive simulation studies demonstrate that the proposed method outperforms the state-of-the-art methods in estimation accuracy and computational efficiency. The practical merit of our method is demonstrated by an application to the CITE-seq dataset. A flexible implementation of our proposed method is available in the R package \emph{COAP}.

S077: Mordern Distributional Methods for Learning, Inference, and Robustness

Conditional Distribution Testing via Measure Transport

◆ *Yuan Gao, Chenxuan He, Jian Huang, Liping Zhu*

Nankai University, Renmin University of China, The Hong Kong

Polytechnic University, Renmin University of China

Conditional distribution tests assess whether the distribution of responses given covariates is equal across domains. These tests are crucial for verifying the covariate shift assumption in domain adaptation and transfer learning. Covariate distribution shift complicates such tests, and current approaches typically depend on explicit density ratio or conditional density estimation, both of which become computationally and statistically prohibitive with high-dimensional covariates or multivariate responses. We propose a framework that formulates the conditional distribution test as an independence test via auxiliary random vectors and measure transport functions, where the functions are learned using continuous normalizing flows. After transforming the problem into an independence test, we apply the improved projection correlation to measure independence, avoiding explicit density modeling and delivering a scalable procedure. Theoretical guarantees are established through asymptotic analysis with data-splitting techniques. Our approach naturally extends to the K-sample setting. Experiments on both synthetic and real-world data demonstrate robust performance compared with popular alternatives.

Transfer Learning through Enhanced Sufficient Representation: Enriching Source Domain Knowledge with Target Data

◆ *Yeheng Ge, Xueyu Zhou, Jian Huang*

The Hong Kong Polytechnic University, The Hong Kong Polytechnic University, The Hong Kong Polytechnic University

Transfer learning is an important approach for addressing the challenges posed by limited data availability in various applications. It accomplishes this by transferring knowledge from well-established source domains to a less familiar target domain. However, traditional transfer learning methods often face difficulties due to rigid model assumptions and the need for a high degree of similarity between source and target domain models. In this paper, we introduce a novel method for transfer learning called Transfer learning through Enhanced Sufficient Representation (TESR). Our approach begins by estimating a sufficient and invariant representation from the source domains. This representation is then enhanced with an independent component derived from the target data, ensuring that it is sufficient for the target domain and adaptable to its specific characteristics. A notable advantage of TESR is that it does not rely on assuming similar model structures across different tasks. For example, the source domain models can be regression models, while the target domain task can be classification. This flexibility makes TESR applicable to a wide range of supervised learning problems. We explore the theoretical properties of TESR and validate its performance through simulation studies and real-world data applications, demonstrating its effectiveness in finite sample settings.

Wasserstein Distributionally Robust Nonparametric Regression

◆ *Changyu Liu, Yuling Jiao, Junhui Wang, Jian Huang*

Wuhan University, Wuhan University, The Chinese University of Hong Kong, The Hong Kong Polytechnic University

Wasserstein distributionally robust optimization (WDRO) strengthens statistical learning under model uncertainty by minimizing the local worst-case risk within a prescribed ambiguity set. Although WDRO has been extensively studied in parametric settings, its theoretical properties in nonparametric frameworks remain underexplored. This paper investigates WDRO for nonparametric regression. We first establish a structural distinction based on the order k of the Wasserstein distance, showing that k=1 induces Lipschitz-type regularization, whereas k>1 corresponds to gradient-norm regularization. To address model misspecification, we analyze the excess local worst-case risk, deriving non-asymptotic error bounds for estimators constructed using norm-constrained feedforward neural networks. This analysis is supported by new covering number and approximation bounds that simultaneously control both the function and its gradient. The proposed estimator achieves a convergence rate of n^{-2\beta/(d+2\beta)} up to logarithmic factors, where \beta depends on the target's smoothness and network parameters. This rate is shown to be minimax optimal under conditions commonly satisfied in high-dimensional settings. Moreover, these bounds on the excess local worst-case risk imply guarantees on the excess natural risk, ensuring robustness against any distribution within the ambiguity set. We show the framework's generality across regression and classification problems. Simulation studies and an application to the MNIST dataset further illustrate the estimator's robustness.

AI-Driven Personalization: Causal Learning + RL for Dynamic Treatment Decisions

Zengjing Chen, Jianqi Feng, ◆ *Lu Wang, Chengzhi Xing*

Shandong University, Shandong University, The Hong Kong Polytechnic University, Zhongtai Securities

In this talk, we present recent advances and statistical causal learning developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. We will first present a tree-based doubly robust reinforcement learning (T-RL) method, which builds a decision tree that maintains the nature of batch-mode reinforcement learning, and then a Stochastic-Tree Search method called ST-RL for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances even with a large number

of covariates. In addition, we consider a common challenge with practical "restrictions" and develop a Restricted Tree-based Reinforcement Learning (RT-RL) method to address this challenge. We illustrate the method using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

S081: New Developments in Survey Sampling

Imperfect extreme ranked set sampling design: A framework for robust and efficient inference

Shuangxi Liu, ◆ *Wangxue Chen, A. M. Elsawah*

Jishou University, Jishou University, Beijing Normal-Hong Kong Baptist University

Extreme ranked set sampling design (ERSSD) is a cost-effective technique for population mean estimation, measuring only extreme units to improve effciency. However, its theoretical advantages rely on the often-unrealistic assumption of perfect ranking. Although imperfect RSSD (IRSSD) been studied, the properties and performance of imperfect ERSSD (IERSSD)-where only extremes are measured under judgment error-remain largely unexplored. This paper fills that critical gap by developing a comprehensive framework for IERSSD that formally accounts for ranking inaccuracies.We derive explicit expressions for the mean squared error and variance of the IERSSD mean estimator. For symmetric populations, the estimator is unbiased. Our analysis shows that IERSSD consistently outperforms simple random sampling design (SRSD) for symmetric distributions and for small samples from asymmetric distributions, even under substantial ranking errors. Compared to IRSSD, IERSSD offers comparable efficiency with simpler implementation when ranking quality is poor or sample sizes are small, providing a valuable efficiency-simplicity trade off. Numerical simulations using exponential, Pareto, normal, and heavy-tailed beta distributions support the theory, showing efficiency gains over SRSD of up to 793% in simulations and 75% in a real ecological dataset. We provide practical guidelines and a decision flowchart for choosing between SRSD, IERSSD, and IRSSD based on distribution type, sample size, and expected ranking quality. The proposed framework establishes IERSSD as a practical, robust, and efficient method for real-world inference where perfect judgment ranking is unattainable.

An improved AdRec model for Missing data in Population Census

◆ *Lingli Sun, Tingting Zhang, Wenqiang Du, Shanshan Qin*

Tianjin University of Finance and Economics

Missing data in population census data (CSD) can affect the accuracy of demographic distributions and, in turn, reduce the reliability of census-based policy decisions. Administrative records data (ARD) provide an important supplementary source for addressing missing CSD, but their quality must be carefully assessed before they are used for imputation. This study improves the Administrative Records model (AdRec) by introducing two methodological extensions. First, LightGBM is used to replace logistic regression in both the person-place (PP) model and the household composition (HC) model, allowing the model to better capture nonlinear relationships and feature interactions. Second, Bagging-LVQ is applied to reclassify coherence levels in a data-driven way, reducing the subjectivity of rule-based classification. Using China's 2014 Urban Household Survey (UHS) as the CSD, we simulate ARD and conduct 100 Monte Carlo replications to evaluate the proposed method. Simulation analyses show that the proposed model can more effectively identify high-quality ARD for addressing missing CSD and can produce reasonable distributions of demographic characteristics, such as gender, age, and educational attainment.

Measuring Cheating and Noncompliance in Survey with Sensitive Questions

Qin Wu

South China Normal University

Refusal and providing untruthful answers from respondents are very common in the survey of collecting highly personal or sensitive issues. Accurate data collections and reliable analysis on these issues are desirable. Both randomized and non-randomized response models are proposed to encourage the respondents to provide truthful answers. But some respondents do not believe that these techniques can protect their anonymity. We propose a new non-randomized model in which the cheating and the noncompliance are both measured. Our proposed techniques are empirically shown to yield accurate estimates for both sensitive and noncompliance probabilities. A real example among university students is used to demonstrate our methodologies.

Jackknife Model Averaging for Complex Survey Data

◆ *Huanxue Zhuang, Xianpeng Zong*

Beijing University of Technology, Beijing University of Technology

Model-assisted inference is widely used in survey sampling to improve efficiency by incorporating auxiliary information, but its performance is sensitive to superpopulation model misspecification. To address this limitation, we introduce model averaging into complex survey inference and propose a Jackknife model averaging approach for unequal probability sampling designs. Under the joint randomness induced by the sampling design and the superpopulation model, we establish

the asymptotic optimality of the proposed model averaging estimator and the consistency of the estimated weights. We further develop a model-assisted averaging estimator for the finite population mean and establish its asymptotic normality. The mean squared error of the proposed estimator is derived, and a consistent estimator is provided. Simulation studies and a real data application demonstrate the effectiveness and reliability of the proposed methods.

S083: New Frontiers in Spatial Modeling: Dissimilarity, Extremes, and High-Dimensional Clustering

Integrating Mechanistic Dynamics to Characterize the Causes of Extremes in Environmental Processes

Megha Pailan, ◆ *Likun Zhang, Christopher K. Wikle*

University of Missouri - Columbia, University of Missouri - Columbia, University of Missouri - Columbia

Extreme weather events, such as droughts, floods, heat domes and wildfires, frequently occur in spatial clusters that evolve over time. In the past two decades, spatial extremes modeling has advanced significantly, particularly through block maxima and peaks-over-threshold models. However, these methods often rely on stringent assumptions about the joint tail behavior and exclude the bulk of the data to focus on extremes. Additionally, copula-based approaches often assume spatial stationarity and temporal independence, neglecting the complex dynamic relationships among multiple factors that interact across space and time. This talk will showcase novel dynamic and causal approaches to spatio-temporal extremes and diverse viewpoints to enhancing the practicality and effectiveness of modeling these extremes.

Generalized Chi-Squared Process for Spatial Dissimilarity

◆ *Xiaotian Zheng, Andrew Zammit-Mangion, Noel Cressie*

University of Georgia, University of New South Wales, University of Wollongong

Spatial dependence and spatial dissimilarity are scientific notions that help explain why ecosystems have more or less biodiversity. One way to quantify spatial dissimilarity is to make pairwise comparisons of groups (or ensembles) at various spatial proximities. Statistical inference to answer science questions for which pairwise-dissimilarity are collected will involve a latent spatial process indexed by two spatial locations. However, little work has been done on this type of spatial process indexed by pairs of locations in a spatial domain. Here, we define such a process that we call a generalized chi-squared process (GCP). We first establish that the GCP is a well-defined stochastic process, with spatially varying generalized chi-squared marginal distributions, and that its finite-dimensional distributions are fully characterized by its covariance function and a so-called shift function. We then embed the GCP within a hierarchical model that enables inference on a latent dissimilarity process for noisy and incomplete pairwise-dissimilarity data. The appeal of the modeling framework is demonstrated through a simulation study and its application to data sets of pairwise dissimilarities of both terrestrial and marine species.

Cluster Size Distributions of Discrete Random Fields

◆ *Dan Cheng, John Ginos*

Arizona State University, Arizona State University

We study discrete random fields parameterized on the integer lattice. For a fixed threshold, the excursion set decomposes into connectedcomponents or clusters, whose size, defined as the number of lattice points they contain, are random. We investigate the probability distribution ofthese cluster sizes. For stationary random fields, we derive exact expressions for the cluster size distribution. To address nonstationary settings, weintroduce a peak-based cluster size distribution, which characterizes the distribution of cluster sizes conditional on the presence of a local maximumabove the threshold. This formulation provides a tractable alternative when exact cluster size distributions are analytically inaccessible. Theproposed framework applies broadly to Gaussian and non-Gaussian random fields, relying only on their joint dependence structure. Our resultsprovide a theoretical foundation for quantifying spatial extent in discretely sampled data, with applications to medical imaging, geoscience,environmental monitoring, and other scientific areas where thresholded random fields naturally arise.

BNPSpace: Bayesian Nonparametric Clustering for Spatially Resolved Transcriptomic Data

Guanyu Hu

Michigan State University

The advent of next-generation sequencing-based spatially resolved transcriptomics (SRT) techniques has reshaped genomic studies by enabling high-throughput gene expression profiling while preserving spatial and morphological context. Nevertheless, there are inherent challenges associated with these new high-dimensional spatial data, such as zero-inflation, over-dispersion, and heterogeneity. These challenges pose obstacles to effective clustering, which is a fundamental problem in SRT data analysis. Current computational approaches often rely on heuristic data preprocessing and arbitrary cluster number prespecification, leading to considerable information loss and consequently, suboptimal downstream analysis. In response to these challenges, we introduce BNPSpace, a novel Bayesian nonparametric spatial clustering framework that directly models SRT count data. BNPSpace facilitates the partitioning of the

whole spatial domain, which is characterized by substantial heterogeneity, into homogeneous spatial domains with similar molecular characteristics while identifying a parsimonious set of discriminating genes among different spatial domains. Moreover, BNPSpace incorporates spatial information through a Markov random field prior model, encouraging a smooth and biologically meaningful partition pattern.

S087: Post and Semi-Supervised Inference

Algorithm-Agnostic Post-Clustering Differential Expression Analysis with Asymptotic FDR Control

Jiajun Liu, ◆ *Mengtao Wen, Changliang Zou, Liu Liu*

Nankai University, Nankai University, Nankai University, Chengdu University of Technology

Post-clustering differential expression (DE) analysis faces the challenge of selection bias known as double-dipping. We present a unified, algorithm-agnostic framework to identify DE variables with asymptotic false discovery rate (FDR) control. This framework consists of: (i) a population-level DE measure defined by the increase in risk after cyclically permuting cluster parameters; (ii) a cross-fitting estimator that avoids degenerate null distributions and achieves asymptotic normality under mild conditions on the clustering excess risk, thereby enabling valid inference even when cluster parameters converge slowly; (iii) a p-value-free multiple testing procedure that employs a "resplit-per-variable" strategy to mitigate dependence among null statistics and incorporates data aggregation to enhance power. The effectiveness of the proposed framework is demonstrated through extensive simulations and an analysis of singlecell RNA sequencing data.

Shape-Adaptive Conditional Calibration for Conformal Prediction via Minimax Optimization

◆ *Yajie Bao, Chuchen Zhang, Zhaojun Wang, Haojie Ren, Changliang Zou*

Nankai University, Nankai University, Nankai University, Shanghai Jiao Tong University, Nankai University

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 during the calibration phase, rather than simply calibrating a fixed sublevel set. 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. The MOPI 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.

Design and Modeling for Order-of-Addition Factorial Experiments with Interaction Effects

Jianbin Chen

Beijing Institute of Technology

In scientific research and industrial engineering, the response of an experiment is often determined by both the level combinations of factors and the addition sequence. Such experiments are referred to as order-of-addition factorial (OofAF) experiments. The interactions between factor levels and their order of addition frequently have a substantial impact on the response. For instance, in automotive manufacturing, painting components prior to welding may result in heat-induced damage due to welding temperatures, thereby increasing the defect rate. To address these interactions in OofAF experiments, we introduce a class of conditional main effects (CMEs) to capture interaction effects by measuring the conditional effect of a pairwise ordering factor at a fixed level of another factor. Leveraging these CMEs, we propose a statistically interpretable model, called the conditional main effects pairwise ordering model, for analyzing OofAF experiments. Under this model, we develop a systematic construction method to generate a D-optimal fractional OofAF design, which reduces experimental costs while enabling efficient estimation of main effects and CMEs. Comparative case studies are presented to illustrate that this proposed model outperforms existing approaches in predictive accuracy when analyzing OofAF experiments with significant interaction effects.

Semi-Supervised Inference for U-Estimation

Liuhua Peng

University of Melbourne

Incomplete labeled data are ubiquitous across diverse domains, where obtaining fully labeled samples is prohibitively expensive, time-consuming, or privacy-intrusive. This has spurred growing interest in methods that can effectively leverage unlabeled observations. U-estimation, empirical risk minimization based on U-statistics, is widely adopted in regression, classification, ranking, and related problems. We propose a semi-supervised U-estimation enhanced by the abundance of unlabeled data, and

investigate its statistical properties. When the working model is misspecified, the proposed estimator achieves higher efficiency than its supervised counterpart that uses only labeled data; when the model is correctly specified, the two estimators are asymptotically of equal efficiency. Simulation studies and real data analysis confirm the effectiveness of our approach.

S090: Recent Advance in Statistical Learning

False Discovery Rate Control via Data Splitting for Testing-after-Clustering

◆ *Lijun Wang, Yingxin Lin, Hongyu Zhao*

Zhejiang University, The Chinese University of Hong Kong, Yale University

Testing for differences in features between clusters in various applications often leads to inflated false positives when practitioners use the same dataset to identify clusters and then test features, an issue commonly known as ``double dipping''. To address this challenge, inspired by data-splitting strategies for controlling the false discovery rate (FDR) in regressions (Dai et al., 2023), we present a novel method that applies data-splitting to control FDR while maintaining high power in unsupervised clustering. We first divide the dataset into two halves, then apply the conventional testing-after-clustering procedure to each half separately and combine the resulting test statistics to form a new statistic for each feature. The new statistic can control the FDR due to its property of having a sampling distribution that is symmetric around zero for any null feature. To further enhance stability and power, we suggest multiple data splitting, which involves repeatedly splitting the data and combining results. Our proposed data-splitting methods are mathematically proven to asymptotically control FDR in Gaussian settings. Through extensive simulations and analyses of single-cell RNA sequencing (scRNA-seq) datasets, we demonstrate that the data-splitting methods are easy to implement, adaptable to existing single-cell data analysis pipelines, and often outperform other approaches when dealing with weak signals and high correlations among features.

Robust Bayesian Inference on Riemannian Submanifold

◆ *Rong Tang, Anirban Bhattacharya, Debdeep Pati, Yun Yang*

Hong Kong University of Science and Technology

Manifold-valued parameters routinely arise in modern statistical applications such as in medical imaging, robotics, and computer vision, to name a few. While traditional Bayesian approaches are applicable to such settings by considering an ambient Euclidean space as the parameter space, we demonstrate the benefits of integrating manifold structure into the Bayesian framework, both theoretically and computationally. Moreover, existing Bayesian approaches which are designed specifically for manifold-valued parameters are primarily model-based, which are typically subject to inaccurate uncertainty quantification under model misspecification. In this article, we propose a robust model-free Bayesian inference for parameters defined on a Riemannian submanifold, which is shown to provide valid uncertainty quantification from a frequentist perspective. Computationally, we propose a Markov chain Monte Carlo to sample from the posterior on the Riemannian submanifold, where the mixing time, in the large sample regime, is shown to depend only on the intrinsic dimension of the parameter space instead of the potentially muchlarger ambient dimension. Our numerical results demonstrate the effectiveness of our approach on a variety of problems, such as multiple quantile regression, reduced-rank regression, and Fréchet mean estimation.

Estimation and Inference for Nonparametric Expected Shortfall Regression over RKHS

Yue Wang

University of Science and Technology of China

Expected shortfall (ES) has emerged as an important metric for characterizing the tail behavior of a random outcome, specifically associated with rarer events that entail severe consequences. In climate science, the threats of flooding and heatwaves loom large, impacting natural environments and human communities. In actuarial studies, a key observation in modeling insurance claim sizes is that features exhibit distinct effects in explaining small and large claims. This article concerns nonparametric expected shortfall regression as a class of statistical methods for tail learning. These methods directly target upper/lower tail averages and will empower practitioners to address complex questions that are beyond the reach of mean regression-based approaches. Using kernel ridge regression, we introduce a two-step nonparametric ES estimatorthat involves a plugged-in quantile function estimate without sample-splitting. We provide non-asymptotic estimation and Gaussian approximation error bounds, depending explicitly on the effective dimension, sample size, regularization parameters, and quantile estimation error. To construct pointwise confidence bands, we propose a fast multiplier bootstrap procedure and establish its validity. We demonstrate the finite-sample performance of the proposed methods through numerical experiments and an empirical study aimed at examining the heterogeneous effects of different air pollutants and meteorological factors on average and high PM2.5 concentration. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

S091: Recent Advances for Statistical Analysis of Censored Data

Transfer Learning for High-Dimensional Semiparametric Regression with Censored Covariates

Huiqiong Li

Yunnan University

Data scarcity and incomplete covariate information are common in applications such as biomedicine, biostatistics, and reliability studies, where collecting large-scale, fully observed target-domain data is often difficult. Existing transfer learning methods typically rely on distributional similarity between source and target domains and require complete target-domain data, limiting their applicability to semiparametric models with censored covariates. This paper develops a transfer learning framework for high-dimensional semiparametric models with censored covariates. To account for censoring, we propose a data-adaptive weighting procedure based on stabilized inverse probability weighting across multiple sources. The transfer procedure combines representation transfer and distance-based transfer: deep neural networks are used to approximate the nonparametric component, while distance-based regularization is imposed on the linear component. The framework allows for several types of source-target heterogeneity, including covariate shift, residual shift, and parameter shift. We establish identifiability conditions for the target model, derive convergence rates for the primary target parameters, and provide theoretical guarantees for transfer-source detection. We further develop an orthogonal debiasing procedure for inference in the target model that incorporates transferable information from auxiliary sources. Simulation studies show that the proposed method improves estimation and inference under limited target-domain sample sizes and heterogeneous source domains. An application to anticancer drug response prediction illustrates its empirical performance and practical relevance.

Variable Selection for Interval-Censored Data with Informative Observation Times

◆ *Ni Li, Mengxin Wang*

Hainan Normal University, Hainan Normal University

In this paper, we consider a general class of failure time data known as case K interval-censored data, which includes case I and case II interval-censored data as special cases. We propose a unified penalized variable selection procedure based on the proportional hazards (PH) model. The method minimizes a negative sieve log-likelihood function combined with a broken adaptive ridge penalty, whereinitial values are obtained from the ridge regression estimator. The proposed approach accommodates informative observation times through conditional modeling that incorporates the history or filtration of the observation process-a scenario that occurs frequently and, if ignored, may lead to biased or misleading estimates. We establish that the proposed selection method possesses the oracle property and that the resulting estimator is semiparametrically efficient. Numerical studies indicate that the proposed approach performs well in practical settings. Finally, the method is applied to a diabetic retinopathy study that motivated this investigation.

Weighted Jackknife Model Averaging Method for Right Censored Data Under Varying Coefficient Models

Binxia Liu

Taiyuan Normal University

In this paper, we consider a novel weighted model averaging method for right censored data under varying-coefficient models. The weights of candidate models are obtained by maximizing a weighted Jackknife weight choice criterion involving the B-spline, weighted least square and Kaplan-Meier estimation techniques. Under some regularity conditions, the proposed model averaging estimator is shown to be asymptotic optimality. Some simulation studies are conducted to evaluate the performace of the proposed model averaging method among other exsiting model selection and model averaging method, which show that our proposed method is always superior to others. Finally, a real data is provided to illustrate the proposed method.

Distributed Sparse Learning for Cox Models Over Networks: A Decentralized Approach

◆ *Yu Zhang, Hui Zhao*

Zhongnan University of Economics and Law, Zhongnan University of Economics and Law

The rapid growth and increasing heterogeneity of data have driven a shift from centralized to decentralized learning frameworks, which offer improved scalability and privacy through peer-to-peer communication. However, sparse Cox regression in decentralized settings remains challenging due to the global dependency of risk sets and heterogeneous censoring mechanisms across sites. In this paper, we propose deCOX, a novel decentralized algorithm for sparse Cox regression under heterogeneity. To accommodate varying baseline hazards, we adopt a stratified Cox model that defines risk sets locally at each node. To mitigate cross-node information inconsistency induced by heterogeneous censoring, we develop a finite-round neighborhood consensus scheme that aggregates local gradients and Hessians into a tractable surrogate least squares objective. We provide a rigorous theoretical analysis of the error induced by consensus iterations and second-order approximation, and show that deCOX achieves linear convergence to the global optimum. Moreover, the proposed estimator is statistically consistent and recovers the true sparse support. Extensive simulations and a real breast cancer survival

study demonstrate that deCOX is robust to heterogeneity and network structures, outperforming existing distributed survival methods in both accuracy and efficiency.

S093: Recent Advances in Complex Data Analysis and New Applications

Slacked Empirical Likelihoods for Post-Criterion Inference

Yiyuan She

Westlake University

Statistical inference under nonsmooth penalties poses significant challenges for traditional empirical likelihood methods, particularly outside low-dimensional settings. We introduce Slacked Empirical Likelihood (SEL), derived from the penalized criterion itself. By utilizing slack variables, SEL converts rigid structural conditions into a tractable dual formulation over a general family of divergence functions. This approach makes penalty-induced structural bias explicit, providing a transparent distinction between noncentral and central limiting behaviors. Through a novel dual geo-centering scheme, SEL naturally absorbs the bias energy to recover a central limit distribution. Our theoretical framework spans both classical and high-dimensional regimes, establishing sample-complexity bounds that capture the trade-offs between data heavy-tailedness, active support size, and non-vanishing regularizer bias.

Geometric standardized mean difference and its application to meta-analysis

Jiandong Shi, Xiaochen Zhang, Lu Lin, Hiu Yee Kwan, ◆ *Tiejun Tong*

Chinese University of Hong Kong, Beijing Normal University at Zhuhai, Shandong University, Hong Kong Baptist University, Hong Kong Baptist University

The standardized mean difference (SMD) is a widely used measure of effect size, particularly common in psychology, clinical trials, and meta-analysis involving continuous outcomes. Traditionally, under the equal variance assumption, the SMD is defined as the mean difference divided by a common standard deviation. This approach is prevalent in meta-analysis but can be overly restrictive in clinical practice. To accommodate unequal variances, the conventional method averages the two variances arithmetically, which does not allow for an unbiased estimation of the SMD. Inspired by this, we propose a geometric approach to averaging the variances, resulting in a novel measure for standardizing the mean difference with unequal variances. We further propose the Cohen-type and Hedges-type estimators for the new SMD, and derive their statistical properties including the confidence intervals. Simulation results show that the Hedges-type estimator performs optimally across various scenarios, demonstrating lower bias, lower mean squared error, and improved coverage probability. A real-world meta-analysis also illustrates that our new SMD and its estimators provide valuable insights to the existing literature and can be highly recommended for practical use.

Time-Varying High Quantile Estimation for Nonstationary Tail Dependent Time Series

◆ *Ting Zhang, Yu Shao*

University of Georgia, Ernst & Young

We consider the problem of time-varying high quantile estimation for a general class of nonstationary tail dependent time series. Unlike the conventional quantile analysis in which the quantile level is typically assumed to be fixed, the current high quantile setting requires the quantile level to grow with the sample size to capture the tail phenomenon. In addition, for nonstationary time series data, the underlying data generating mechanism is expected to change over time, and as a result the associated time-varying high quantile can no longer be treated as a constant parameter making existing results not directly applicable. In this talk, we model the underlying time-varying high quantile as a nonparametric function of time to avoid any parametric misspecification, and we establish the consistency and central limit theorem of nonparametric high quantile estimators for a general class of nonstationary tail dependent time series. It can be seen from our limit theorems that the asymptotic behavior of nonparametric high quantile estimators can be affected by both the double asymptotic scheme caused by the growing quantile level and the serial tail dependence in time series data. Numerical results are provided to illustrate the developed results, and an extension to time-varying high quantile regression models is also discussed.

Model Estimation and Variable Selection for Longitudinal Partial Linear EV Models with Measurement Errors in All Covariates

Mingtao Zhao

Anhui University of Finance and Economics

In this talk, we consider model estimation and variable selection for longitudinal partial linear errors-in-variables (EV) models with measurement errors in both parametric and nonparametric covariates. A bias-corrected penalized empirical likelihood method is proposed using empirical likelihood and quadratic inference functions. The proposed method can simultaneously handle unknown within-subject correlations and measurement errors, and is capable of conducting model estimation and variable selection simultaneously. Under some regularity conditions,

the resulting estimators are shown to be asymptotically normal and possess the oracle property. Numerical results, including simulation studies and a real example analysis, demonstrate that the proposed method is meaningful in finite samples, which validates its theoretical properties and practical applicability.

S098: Recent Advances in Designs and Statistical Methods for Clinical Trials Research

Statistical Inference for Heterogeneous Treatment Effect with Right-Censored Data from Synthesizing Randomized Clinical Trials and Real-World Data

◆ *Guangcai Mao, Shu Yang, Xiaofei Wang*

Duke University, North Carolina State University, Duke University

The heterogeneous treatment effect plays a crucial role in precision medicine. There is evidence that real-world data, even subject to biases, can be employed as supplementary evidence for randomized clinical trials to improve the statistical efficiency of the heterogeneous treatment effect estimation. In this paper, for survival data with right censoring, we consider estimating the heterogeneous treatment effect, defined as the difference of the treatment- specific conditional restricted mean survival times given covariates, by synthesizing evidence from randomized clinical trials and the real-world data with possible biases. We define an omnibus bias function to characterize the effect of biases caused by unmeasured confounders, censoring, and outcome heterogeneity, and further, identify it by combining the trial and real-world data. We propose a penalized sieve method to estimate the heterogeneous treatment effect and the bias function. We further study the theoretical properties of the proposed integrative estimators based on the theory of reproducing kernel Hilbert space and empirical process. The proposed methodology outperforms the approach solely based on the trial data through simulation studies and an integrative analysis of the data from a randomized trial and a real-world registry on early-stage non-small-cell lung cancer.

Win Ratio Under Bivariate Mixed Censored Data Based on Death Time and Progression Time

◆ *Yuan Wu, Xiaofei Wang, Zhiguo Li*

Duke University, Duke University, Duke University

The win-ratio is a powerful framework for prioritized multiple time-to-event outcomes, yet its application has been largely limited to right-censored data. In clinical settings like oncology, intermediate events such as disease progression are often interval-censored, rendering traditional pairwise comparisons infeasible due to data coarsening beyond right-censoing. We propose a novel spline-sieve plug-in estimator for the win-ratio that accommodates a practical bivariate structure: mixed-censoring (one right- and one interval-censored event). We establish the asymptotic normality of our estimator and demonstrate robust finite-sample performance through extensive simulations. Finally, we demonstrate the clinical utility of our method through a real-world application, offering a robust framework for treatment evaluation with complex data structures.

Deconfounded-Debiased Distributed Estimation for High-Dimensional Multi-Site Data with Heterogeneous Pervasive Hidden Confounders

Guoyou Qin

Fudan University

In observational studies, high-dimensional multi-site data are subject to heterogeneous pervasive hidden confounders across sites, leading to substantial bias in distributed estimation. To address this challenge, we propose a deconfounded-debiased distributed estimation method, which integrates majority voting for variable selection and the privacy-preserving aggregation of local deconfounded-debiased estimators at the central site. This approach corrects biases from both heterogeneous pervasive hidden confounders and high-dimensional estimation, while also accommodating site-specific heteroscedasticity in the random errors. Theoretically, we prove that local deconfounded-debiased estimators are asymptotically normal, and local individual hypothesis tests are asymptotically valid with an asymptotic lower bound on their power. Furthermore, we establish variable selection consistency and asymptotic normality for the proposed distributed estimator. We also provide finite-sample guarantees for variable selection, including an upper bound on the expected false positive rate and a lower bound on the expected true positive rate. Simulation experiments and an application to a protein dataset from the UK Biobank demonstrate our method's superior finite-sample performance.

S105: Recent Advances in Selective Inference, Data Assimilation, and Functional Data

AI-Driven Personalization: Causal Learning + RL for Dynamic Treatment Decisions

Yezi Wang, ◆ *Lu Wang*

Central South University, Central South University

In this talk, we present recent advances and statistical causal learning developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification

of an optimal DTR is a key component for precision medicine and personalized health care. We will first present a tree-based doubly robust reinforcement learning (T-RL) method, which builds a decision tree that maintains the nature of batch-mode reinforcement learning, and then a Stochastic-Tree Search method called ST-RL for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances even with a large number of covariates. In addition, we consider a common challenge with practical "restrictions" and develop a Restricted Tree-based Reinforcement Learning (RT-RL) method to address this challenge. We illustrate the method using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

Localization Estimator for High Dimensional Tensor Covariance Matrices

◆ *Haoxuan Sun, Song Xi Chen, Yumou Qiu*

Harbin Institute of Technology

This paper considers covariance matrix estimation of tensor data under high dimensionality. A multi-bandable covariance class is established to accommodate the need for complex covariance structures of multi-layer lattices and general covariance decay patterns. We propose a high dimensional covariance localization estimator for tensor data, which regulates the sample covariance matrix through a localization function. The statistical properties of the proposed estimator are studied by deriving the minimax rates of convergence under the spectral and the Frobenius norms. Numerical experiments and real data analysis on ocean eddy data are carried out to illustrate the utility of the proposed method in practice.

A Unified Framework for Large-Scale Inference of Classification: Error Rate Control and Optimality

◆ *Yinrui Sun, Yin Xia*

Fudan University, Fudan University

Classification is a fundamental task in supervised learning, while achieving valid misclassification rate control remains challenging due to possibly the limited predictive capability of the classifiers or the intrinsic complexity of the classification task. In this talk, we address large-scale multi-class classification problems with general error rate guarantees to enhance algorithmic trustworthiness. To this end, we first introduce a notion of group-wise classification, which unifies the common class-wise and overall classifications as special cases. We then develop a unified inference framework for the general group-wise classification that consists of three steps: Pre-classification, Selective p-value construction, and large-scale Post-classification decisions (PSP). Theoretically, PSP is distribution-free and provides valid finite-sample guarantees for controlling general group-wise false decision rates at target levels. To show the power of PSP, we demonstrate that the step of post-classification decisions never degrades the power of pre-classification, provided that pre-classification has been sufficiently powerful to meet the target error levels. We further establish general power optimality theories for PSP from both non-asymptotic and asymptotic perspectives. Numerical results in both simulations and real data analysis validate the performance of the proposed PSP approach.

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

Rongyi Sun, Wenguang Sun, ◆ *Zinan Zhao*

Zhejiang University, Zhejiang University, Zhejiang University

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, rendering 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 patterns. 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 assumption of pairwise exchangeability, offering guaranteed error rate control, enhanced statistical power, and improved interpretability in detecting structured anomalies.

S111: Recent Advances in Statistical Methods for Genetic Data

Robust Multi-Ancestry Cis-Mendelian Randomization Method with Bayesian Fine-Mapping

Chengli Zhang, Chong Wu, ◆ *Haoran Xue*

City University of Hong Kong, University of Texas MD Anderson Cancer Center, City University of Hong Kong

Mendelian randomization (MR) has become a powerful tool for inferring causal relationships among human traits using genetic variants (mainly single nucleotide polymorphisms, SNPs) as instrumental variables (IVs). cis-MR exploits correlated SNPs within local genetic regions and is suitable for studying causal effects from molecular exposures (e.g., gene expression levels

and protein levels) to complex traits. However, the widespread pleiotropy of genetic variants may compromise the validity of cis-MR, leading to false discoveries of causal factors. In addition, differences in effect sizes and correlation structures (i.e., linkage disequilibrium, LD) of genetic variants between different ancestries limit the power of cis-MR when analyzing each ancestry separately. To address these challenges, we propose a robust multi-ancestry cis-MR method that employs a Bayesian fine-mapping approach to detect invalid IVs, enabling efficient integration of data from different ancestries. By applying this method to the Atherosclerosis Risk in Communities Study (ARIC) and the Million Veteran Program (MVP) data, we identified putative causal proteins that are specific to or shared between European and African ancestries.

Knockoff-Augmented Meta-Analysis Identifies Shared and Ancestry-Specific Risk Loci in Multi-Ancestry Genome-Wide Association Studies

Jian Yin, ◆ *Yi Yang*

City University of Hong Kong, City University of Hong Kong

The increasing availability of non-European individuals in biobanks has enabled multi-ancestry meta-analysis to improve the discovery of genetic risk loci for complex traits and diseases in genome-wide association studies (GWAS). However, most meta-analysis frameworks yield only aggregate genetic effects under the strong assumption that causal variants are universally shared, disregarding the fact that nearly half of causal variants are ancestry-specific. Moreover, conventional meta-analysis typically controls the family-wise error rate, a highly conservative approach in high-dimensional genetic data that substantially reduces statistical power compared to false discovery rate (FDR) control. Here, we propose KAMA, a knockoff-augmented meta-analysis framework that identifies both shared and cohort-specific risk loci under FDR control while simultaneously pinpointing their exact contributing cohorts. Applying KAMA to summary-level multi-ancestry GWAS data for three complex traits revealed that non-European cohorts drive approximately 50% to 65% of the identified risk loci. We further demonstrate KAMA's versatility across diverse GWAS data types and study designs through a meta-analysis of population- and family-based GWAS for autism spectrum disorder.

A Robust and Computationally Scalable Method for Integrating a Large Number of Genomic Studies

◆ *Zhiang Yuan, Hongyuan Cao, Yingying Wei*

The Chinese University of Hong Kong, Florida State University, The Chinese University of Hong Kong

While individual genomic studies can suffer from weak signals, genomic studies spanning diverse conditions are becoming increasingly prevalent. For example, the integration of multi-ethnic genome-wide association studies is now essential to bolster power in underrepresented minority cohorts. Meanwhile, technologies like Perturb-seq have enabled genome-scale screens of transcriptional responses across hundreds of knockouts. However, conducting meta-analysis across numerous conditions presents significant challenges: the total number of association patterns 2^D grows exponentially with the condition number D. Moreover, existing meta-analysis methods typically rely on parametric assumptions that are not applicable to Perturb-seq studies, which usually adopt nonparametric tests for differential expression. Here, we propose a novel method that is able to (a) detect condition-specific signals while borrowing strength across conditions, (b) scale to a large number of conditions and (c) estimate non-null p-value distributions nonparametrically. We establish model identifiability and asymptotic false discovery rate control. Numerical studies show that our method outperforms existing methods.

S119: Regression Modeling and Efficient Estimation for Complex Data

Bayesian Spatiotemporal Modeling for the Inpatient Hospital Costs of Alcohol-Related Disorders

Zhen Yu

University of International Business and Economics

Understanding how health care costs vary across different demographics and health conditions is essential to developing policies for health care cost reduction. It may not be optimal to apply the conventional mean regression due to its sensitivity to the high level of skewness and spatio-temporal heterogeneity presented in the cost data. To find an alternative method for spatio-temporal analysis with robustness and high estimation efficiency, we combine information across multiple quantiles and propose a Bayesian spatio-temporal weighted composite quantile regression (ST-WCQR) model. An easy-to-implement Gibbs sampling algorithm is provided based on the asymmetric Laplace mixture representation of the error term. Extensive simulation studies show that ST-WCQR outperforms existing methods for skewed error distributions. We apply ST-WCQR to investigate how patients' characteristics affected the inpatient hospital costs for alcohol-related disorders and identify areas that could be targeted for cost reduction in New York State from 2015 to 2017.

Subgroup Identification with Latent Factor Structure

Yong He, Dong Liu, Fuxin Wang, ◆ *Mingjuan Zhang, Wen-Xin Zhou*

Shandong University, National University of Singapore, University of Wisconsin-Madison, Shanghai Lixin University of Accounting and Finance, University of Illinois Chicago

Subgroup analysis has garnered growing attention for its ability to uncover meaningful subgroups within heterogeneous populations, thereby enhancing predictive power. However, in fields such as the social sciences and econometrics, covariates are often highly correlated due to latent common factors. This poses substantial challenges for subgroup identification, a problem largely overlooked in the literature. To address this gap, we propose a center-augmented subgroup identification method within the Factor Augmented (sparse) Linear Model framework. The proposed method accommodates strong dependence among covariates and offers computational advantages, achieving complexity $O(nK)$, compared to the $O(n^2)$ complexity of conventional pairwise fusion penalty methods, where n denotes the sample size and K the number of subgroups. For implementation, we develop two algorithms: a Cyclic Coordinate Descent (CCD) method based on an ℓ_2-type distance, and a Difference-of-Convex Alternating Direction Method of Multipliers (DC-ADMM) based on an ℓ_1-type distance. Under the diverging-dimension regime, we analyze the asymptotic properties of both the oracle estimator and the local minimizer and establish the finite-step convergence of the DC-ADMM algorithm to a Karush–Kuhn–Tucker (KKT) point. The effectiveness of the proposed approach is demonstrated through extensive numerical experiments and a real-data application investigating the impact of demographic structure on economic growth. An accompanying \texttt{R} package, \texttt{SILFS}, is available on CRAN\footnote{\url{https://cran.r-project.org/web/packages/SILFS/index.html}}

A Nested Singular Value Decomposition-Based Surrogate Model for Spatiotemporal Flows

Shixiang Li

ZJUI Institute, Zhejiang University

High-fidelity simulators serve as effective alternatives to physical experiments for understanding complex dynamics, which are often computationally expensive. While surrogate models enable rapid simulations, they typically fail to capture dynamic information from high-dimensional spatiotemporal flow fields. To address this, we propose a tailored multi-output Gaussian process (MOGP) surrogate model for spatiotemporal emulation, incorporating uncertainty quantification. Our approach leverages singular value decomposition to project high-dimensional spatiotemporal data into a low-dimensional latent space, with explicit analysis of decomposition errors. This reformulates the MOGP as a linear combination of independent Gaussian processes, each modeling the relationship between projection coefficients and simulator inputs. The resulting surrogate achieves reliable prediction of unobserved fluid dynamics while significantly accelerating computation, facilitating efficient design exploration. Validation through numerical experiments and real-data case studies demonstrates superior accuracy and efficiency over existing methods.

Collaborative Estimation of Quantile Treatment Effect for Distributed Data

◆ *Ye Fan, Ying Wei, Sung Nok Chiu, Tiejun Tong, Nan Lin*

Capital University of Economics and Business, Columbia University, Hong Kong Baptist University, Hong Kong Baptist University, Washington University in St. Louis

Quantile treatment effect (QTE) provides a powerful tool for capturing treatment effect heterogeneity, which is particularly useful in precision medicine, such as Alzheimer's Disease (AD) research. However, computing QTE for distributed data still remains challenging. Federated learning, a popular technique for collaborative distributed data analysis, suffers from significant efficiency degradation when local data sizes are unbalanced, a scenario frequently encountered in AD research. In this work, we propose a Sequential Collaborative QTE (SCQTE) method for estimating QTE in distributed data, which leverages the renewable estimation and convolution smoothing techniques. SCQTE enables efficient implementation of both conditional and unconditional QTE estimation, with no restrictive requirements on local data sizes in distributed environments. Under mild regularity conditions, we prove that the SCQTE estimator is consistent, asymptotically normal, and asymptotically equivalent to the oracle estimator derived from full data. Additionally, SCQTE exhibits the favorable doubly robust property. It can consistently estimate QTE as long as either the propensity score model or the outcome model is correctly specified. Simulation studies demonstrate that SCQTE achieves estimation accuracy comparable to traditional centralized methods and outperforms federated learning when local data block sizes are small. We further validate the practical utility of SCQTE by analyzing data from the National Alzheimer's Coordinating Center (NACC).

S132: Statistical Inference and Subgroup Analysis for Specialized Data Types

The MLE for Sparse Rasch Model: Consistency, Top-k Recovery and Gaussian Approximation

Pai Peng, ◆ *Lianqiang Qu, Qiuping Wang, Shufang Wang, Ting Yan*

Jianghan University, Central China Normal University, Zhaoqing University, Central China Normal University, Central China Normal University

The Rasch model has been widely used to analyse item response data in psychometrics and educational assessments. When the number of individuals and items are large, it may be impractical to provide all possible responses. It is desirable to study sparse item response experiments. Here, we propose to use the Erd\H{o}s\textendash R\'enyi random sampling design, where an individual responds to an item with low probability p. We prove the uniform consistency of the maximum likelihood estimator when both the number of individuals, n, and the number of items, m, approach infinity. Sampling probability p can be as small as $\max\{\log^\delta(m)/n, \log^\delta(n)/m\}$ up to a constant factor with δ is a constant, which is a fundamental requirement to ensure the connection of the sampling graph by the theory of the Erd\H{o}s\textendash R\'enyi graph. The key technique behind this significant advancement is a powerful

leave-one-out method for the Rasch model. We further establish the asymptotical normality of the MLE by using a simple matrix to approximate the inverse of the Fisher information matrix. The theoretical results are corroborated by simulation studies and an analysis of a large item-response dataset.

Conditional Inference for High-Dimensional Multi-Omics Survival Data

Heyuan Zhang, ◆ *Meiling Hao, Lianqiang Qu, Liuquan Sun*

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, University of International Business and Economics, Central China Normal University, University of Chinese Academy of Sciences

Multi-omics data present significant challenges for statistical inference due to the complex interdependencies among biological layers.

In this paper, we introduce a novel Multi-Omics Factor-Adjusted Cox (MOFA-Cox) model for analyzing multi-omics survival data, effectively addressing the intricate correlations across various omics layers.

We provide a factor-adjusted decorrelated score test for the MOFA-Cox model in high-dimensional survival analysis.

Our method accommodates situations where the dimension of the parameters being tested exceeds the sample size,

while not imposing a sparsity assumption on them.

We establish the limiting null distribution of the proposed test and analyze its power under local alternatives.

Numerical studies and an application to the TCGA breast cancer dataset demonstrate the effectiveness of our method.

Large-Scale Quantile Regression with Covariates Missing at Random

◆ *Xiaoran Yang, Yiwei Wang, Fangfang Bai*

Shandong Technology and Business University, University of International Business and Economics, University of International Business and Economics

Large-scale data has become increasingly prevalent across various domains due to modern technological advancements. Analyzing this extremely large volume of data is inherently complex, as it often exhibits high skewness, heteroscedasticity of variance, and missing values. Quantile regression (QR) is a popular method for analyzing heterogeneous data. In this article, we fit a linear quantile regression model for large-scale data with randomly missing covariates. We integrate information from multiple machines by constructing a communication-efficient surrogate loss (CSL) function. This function combines a weighting method to address missing data with a convolution-type smoothing procedure to deal with the non-smooth quantile regression loss function. Under mild conditions, our proposed estimator demonstrates consistency and asymptotic normality. The proposed approach is evaluated through extensive simulation studies and applied to the survey data from the Behavioral Risk Factor Surveillance System (BRFSS).

Online Inference Based on Streaming Time-to-Event Data

◆ *Xin Chen, Hua Liu, Jiaqi Men, Jinhong You*

Shanghai Lixin University of Accounting and Finance

Time-to-event data arise in a wide range of applications, including reliability engineering, clinical prognosis, and large-scale monitoring systems, where observations are collected sequentially and the observed event indicators and event times may evolve over time. Existing online survival analysis methods typically focus on the accummulation of new observations, but treat the observed event times as fixed once observed. To address this limitation, we propose a statistical framework for fully streaming survival data that accommodates not only the sequential arrival of observations but also the ongoing updates of event indicators and times. Our method provides consistent estimates without constraints on batch size and includes online inference tools, such as sequentially updated confidence intervals and a hazard-monitoring algorithm. The procedure is computationally tractable for large-scale streaming data, and its theoretical properties are established.

S137: Statistical Learning for Complex Data

Network Alternating Direction Method of Multipliers for Ultrahigh-Dimensional Decentralised Federated Learning

Wei Dong

Zhengzhou University

Ultrahigh-dimensional data analysis has received great achievement in recent years. When the data are stored in multiple clients and the clients can be connected only with each other through a network structure, the implementation of ultrahigh-dimensional analysis can be numerically challenging or even infeasible. In this work, we study decentralised federated learning for ultrahigh-dimensional data analysis, where the parameters of interest are estimated via a large amount of devices without data sharing by a network structure. In the local machines, each parallel runs gradient ascent to obtain estimators via the sparsity-restricted constrained methods. Also, we obtain a global model by aggregating each machine's information via an alternating direction method of multipliers (ADMM) using a concave pairwise fusion penalty between different machines through a network structure. The proposed method can mitigate privacy risks from traditional machine learning, recover the sparsity and provide estimates of all regression coefficients simultaneously. Under mild conditions, we show the convergence and estimation consistency of our method. The promising performance of the method is supported by both simulated and real data examples.

Transfer Learning for High-dimensional Accelerated Failure Time Models

Jie Wu

Anhui University

Transfer learning is a powerful tool to leverage information from different but related source domain to boost performance of the target task. Despite the substantial developments in various high-dimensional data analysis, survival data are insufficiently accounted by current transfer learning approaches. Random censoring mechanism, the population heterogeneity of multiple datasets and high dimensionality of predictors pose major challenges. How to properly conduct transfer learning procedure onto survival datasets to improve the statistical inference at the target and simultaneously guarantee theoretical properties remain unclear. In this paper, we propose a new transfer learning method for the high-dimensional right-censored survival data under accelerated failure time (AFT) model. The proposed method applies mean imputation techniques to adjust for censoring data, thus facilitating the subsequent pooling and debiasing transfer procedures. We show that under mild conditions, the proposed method achieves lower error bounds over the existing non-transfer learning methods even there exists obstacles from censoring. We also develop an algorithm based on sample splitting to detect informative source datasets and prove that the selected informative sources is consistent with the true one. Extensive simulation results and a real data case concerning TCGA microarray gene expression are reported to support our theoretical claims and to demonstrate the empirical usefulness of the proposed transfer learning method. To the best of our knowledge, this is the first work on transfer learning for high-dimensional accelerated failure time modelling with theoretical guarantees.

Minimum Distance Kernel Estimation Based Robust Perturbation Subsampling for Big Data Linear Regression

Mingqiu Wang

Qufu Normal University

The advantage of the Minimum Distance Kernel Estimation (MDKE) method lies in its robustness to outliers in both the response and covariates, with no reliance on the density function of the response variable. However, when dealing with large-scale data under limited computing resources, the computational cost of full-sample estimation using the MDKE method becomes prohibitive, rendering it impractical for large-scale applications. To overcome this challenge, this paper integrates the MDKE method with the random perturbation subsampling algorithm, introducing two novel subsampling algorithms based on product weights and additive weights, respectively. The proposed MDKE-based method eliminates the need to compute explicit subsampling probabilities for all observations, substantially reducing the computational burden. We further establish the consistency and asymptotic normality of the corresponding estimators, demonstrating that the estimator obtained from the additive weighting algorithm achieves superior robustness. The effectiveness and robustness of the MDKE-based method are evaluated by extensive simulation studies and a real-data application.

A Novel Skinny Variational Bayesian Estimation Method for High-Dimensional Cox Models with Interval-Censored Data

◆ *Wenting Liu, Huiqiong Li, Niansheng Tang*

Xi'an University of Finance and Economics, Yunnan University, Yunnan University

High-dimensional interval-censored survival data are increasingly encountered in fields such as biomedical research,finance, and the social sciences, presenting substantial challenges for variable selection and parameter estimation.In this paper, we introduce a novel Skinny Variational Bayesian (SVB) algorithm designed for efficient parametric inference in high-dimensional Cox proportional hazards models with interval-censored outcomes. The proposed method combines the Spike-and-Slab sparse prior with the computational efficiency of the Skinny Gibbs

sampler, striking an effective balance between scalability and the inferential rigor of Bayesian methodology. Extensive simulation studies demonstrate that the SVB algorithm consistently outperforms existing frequentist and Bayesian methods in terms of variable selection accuracy, estimation precision. The utility of the proposed approach is further validated through application to data from an HIV vaccine trial, where it successfully identifies key immunological biomarkers associated with HIV-1 infection risk and disease progression.

S145: Statistics, Machine Learning, and AI

Nested Deep Learning Model Towards a Foundation Model for Brain Signal Data

Haipeng Shen

University of Hong Kong

Epilepsy affects over 50 million people globally, with EEG/MEG-based spike detection playing a crucial role in diagnosis and treatment. Manual spike identification is time-consuming and requires specialized training, limiting the number of professionals available to analyze EEG/MEG data. To address this, various algorithmic approaches have been developed. However, current methods face challenges in handling varying channel configurations and in identifying the specific channels where spikes originate. This paper introduces a novel Nested Deep Learning (NDL) framework designed to overcome these limitations. NDL applies a weighted combination of signals across all channels, ensuring adaptability to different channel setups, and allows clinicians to identify key channels more accurately. Through theoretical analysis and empirical validation on real EEG/MEG datasets, NDL demonstrates superior accuracy in spike detection and channel localization compared to traditional methods. The results show that NDL improves prediction accuracy, supports cross-modality data integration, and can be fine-tuned for various neurophysiological applications.

DeepSuM: A Deep Sufficient and Efficient Modality Learning Framework

Ting Li

Southern University of Science and Technology

Multimodal learning has become a pivotal approach in developing robust learning models with applications spanning multimedia, robotics, large language models, and healthcare. The efficiency of multimodal systems is a critical concern, given the varying costs and resource demands of different modalities. This underscores the necessity for effective modality selection to balance performance gains against resource expenditures. In this study, we propose a novel framework for modality selection that independently learns the representation of each modality. This approach allows for the assessment of each modality's significance within its unique representation space, enabling the development of tailored encoders and facilitating the joint analysis of modalities with distinct characteristics. Our framework aims to enhance the sufficiency and effectiveness of multimodal learning by optimizing modality integration and selection.

In-Sample Evaluation of Subgroups Identified by Generic Machine Learning

Xinzhou Guo

Hong Kong University of Science and Technology

When a subgroup is identified from the data, it must be evaluated in a replicable way. The usual in-sample approach, which evaluates the post-hoc identified subgroup as predefined, might suffer from selection bias. This issue of in-sample evaluation of data-dependent objects is well recognized but particularly challenging here. Unlike discrete or finite-dimensional data-dependent objects addressed before, the selection bias here is induced by post-hoc identified subgroups, data-dependent sets potentially defined by infinite-dimensional functionals with nonsmooth boundaries known as nonregularity. The out-of-sample approach, which splits data for subgroup identification and evaluation, can help address selection bias but might suffer from efficiency loss and instability. In this paper, we propose a conditional adaptive perturbation approach to remove selection bias in in-sample subgroup evaluation and deliver valid inference on subgroups identified from the whole dataset by generic machine learning, regardless of whether regularity is satisfied. The proposed method is easy-to-compute, allows model-free and even black-box subgroup identification, and achieves full efficiency across broad scenarios of subgroup analysis through a novel theoretical framework of triple robustness linking rates of subgroup identification and nuisance estimation. The merits of the proposed method are demonstrated by a re-analysis of the ACTG 175 trial.

Signed Network Embedding: Balance, Community and Heterogeneity

Junhui Wang

Chinese University of Hong Kong

Network data has attracted growing interest across scientific domains, prompting the development of various network models. Existing network analysis methods mainly focus on unsigned networks, whereas signed networks, consisting of both positive and negative edges, have been frequently encountered in practice but much less investigated. In this talk, we will formally define strong and weak balance in signed networks, and propose a

signed network embedding model (SNEM), which is capable of modeling strong- and weak-balanced signed networks simultaneously. We will develop an efficient two-step algorithm to maximize the resulting likelihood function and recover the community membership. The effectiveness of SNEM is supported by its asymptotic properties in terms of both community detection and probability estimation as well as its superior numerical performance on a variety of synthetic and real networks.

S057: Innovative Machine Learning and Statistical Methods for Complex Biomedical Data

Multi-Modal Representation Learning with Partially Pairwise Observations

Younggeun Kim

Ulsan National Institute of Science and Technology

Biomedical data are frequently in a multi-modal structure, e.g., structural and functional neuroimaging accompanied by clinical covariates or genomics. Such multiple modalities give opportunities to deeper understanding underlying patterns in disorders of interest; however, they pose challenges in addressing complicated between-modality dependencies and increased missing modalities. In this talk, we propose a novel representation learning method for partially-observed modalities. Our method combines several identifiable generative models so that, even with missed modalities, the remaining partially observed modalities can identify dependency structures between observed ones, and by integrating these findings, ultimately disentangle all of joint information and modality-specific information without full modality observations.

Evaluating Multiplex Diagnostic Test Using Multivariable Isotonic Regression

◆ *Ying Kuen Cheung, Louise Kuhn*

Columbia University, Columbia University

In a diagnostic test using multiplex assay, each individual biomarker is often expected to have monotonic association with the disease outcome, and, therefore, the underlying disease classification rule is partially ordered with respect to the biomarkers. Nonparametric estimation of the classification rule can be accomplished by projecting an unconstrained Bayes estimator onto the partial ordering subspace. However, computing the projection is challenging as it involves performing maximization over a constrained parameter space whose size grows exponentially with the sample size. We introduce a novel sequential update method for projection-based nonparametric estimation of the disease classification rule and propose new recursive algorithms to implement the method. The proposed algorithms yields the exact Bayes solution that maximizes the posterior gain with respect to a classification-type gain function. When compared to an existing algorithm that gives approximate Bayes solution, our algorithms accomplish the same tasks with much reduced elapsed time in simulation study. We apply the sequential update method to evaluate a human papillomavirus test for cervical cancer precursor lesions and derive diagnostic rule that improves accuracy on existing estimation methods including parametric logistic regression and monotone generalized additive models.

MACHINE: A Robust Multi-Ancestry Fine-Mapping Method Using a Continuous Shrinkage Prior

Xiang Li, Zewei Xiong, Pak Chung Sham, ◆ *Yan Dora Zhang*

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

Fine-mapping seeks to identify causal variants from GWAS signals. Multi-ancestry data can improve resolution by leveraging differences in linkage disequilibrium, but existing methods often assume that causal variants are fully shared across ancestries and rely on discrete priors. This presentation introduces MACHINE, a Bayesian multi-ancestry fine-mapping framework that uses a continuous global–local shrinkage prior to model both shared and ancestry-specific effects. MACHINE also enables false discovery rate control with GWAS summary statistics and out-of-sample LD matrices, and it can incorporate functional annotations through g-LDSC to improve performance. Simulations and real-data analyses of lipid traits and schizophrenia show that MACHINE + g-LDSC is robust, achieves better FDR control than existing methods, and identifies novel causal variants across ancestries.

Dynamic Progression Model with Individualized Longitudinal Registration

◆ *Zexi Cai, Yuanjia Wang*

University of Macao

Ordinary differential equation (ODE) models have been widely utilized to model dynamic systems, particularly in characterizing longitudinal disease progression. However, existing parameter estimation methodologies frequently overlook crucial sources of individual heterogeneity—such as varying baseline states, intrinsic progression rates, and the impact of time-varying external factors. Ignoring these elements often leads to unstable estimations and substantial biases under finite observational horizons. To address these critical limitations simultaneously, we propose a novel dynamic progression modeling framework integrated with an individualized longitudinal registration mechanism. Our method maps observational data to an underlying internal temporal domain, structurally accommodating subject-specific initial conditions alongside time-varying covariate

effects. By systematically addressing these multiple dimensions of heterogeneity, the proposed framework offers a flexible and robust solution for delineating complex longitudinal trajectories. We establish the asymptotic properties of the proposed estimators and demonstrate the superior numerical stability and practical utility of our approach through extensive simulation studies and a real-world data application.

S151: Time Series Econometrics and Applications

Robust Nonparametric Testing for Structural Changes in Multivariate Volatility via Multiple Quantiles

Jilin Wu

Xiamen University

This paper proposes a nonparametric test for structural changes in multivariate volatility via quantile regression, specifically designed for heavy-tailed time series. By constructing a weighted U-statistic based on quantile residuals of the null hypothesis, our approach circumvents the restrictive moment conditions required by traditional least-squares or quasi-likelihood methods. We establish the consistency of the test against a broad class of alternatives, including both abrupt breakpoints and gradual structural changes. Additionally, we derive the asymptotic power properties under two distinct types of local alternatives. Simulation experiments demonstrate that the proposed test maintains reasonable size control and offers substantial power gains over existing benchmarks in heavy-tailed settings. An application to testing the stability of the covariance matrix of the Fama–French three factors provides new evidence of significant time variation, illustrating the practical utility of our approach.

A Testing for Symmetry via Quantile Regression Martingalization

◆*Jinfeng Zhang, Shaoxin Hong*

Shandong university, Boston college

This paper studies statistical test for symmetry based on the regression quantile process. Asymptotic properties of Kolmogorov-Smirnov statistic are investigated. Testing procedures reproposed based on regression quantiles and martingale transformations are applied to remove the nuisance parameters in the limiting distributions. A simulation study indicates that the proposed test maintains the nominal level of significance fairly accurately even for finite samples, and a comparison with the classical symmetry tests shows that our test has good power for detecting different asymmetric distributions.

Estimating Temperature-Electricity Response Curves: A Mixed-Frequency Semiparametric Panel Data Model

Kerui Du

Xiamen University

Estimating non-linear relationships with mixed-frequency data is a key challenge in applied economics. The common practice of pre-aggregating high-frequency regressors induces significant bias via Jensen's inequality, masking the impact of extreme events. This paper develops a mixed-frequency semiparametric panel data model to address this issue. Using a sieve estimator within a fixed-effects framework, our approach flexibly approximates the unknown response function with the full set of high-frequency observations, thereby avoiding aggregation bias. Monte Carlo simulations demonstrate our estimator substantially outperforms conventional methods, such as the binned estimator, in recovering responses to extremes. Applying the model to China's electricity demand, we find that conventional methods significantly underestimate consumption during periods of extreme heat and cold. This has crucial implications for energy and climate policy. The framework provides a general tool for assessing the impact of high-frequency event profiles on low-frequency outcomes in fields such as finance, labor, and public health.

Estimating the Number of Groups in Nonstationary Panels: A Penalty-Free Perspective

Ruiqi Liu, ◆*Shaoxin Hong*

Texas Tech University, Xiamen University

This paper considers nonstationary panel models with latent group structures when the number of groups is unknown. We introduce a novel approach for estimating the number of groups, employing two innovatively designed within-group dissimilarity measurements. By maximizing the dissimilarity ratio, we develop two estimators that do not rely on tuning parameters, simplifying practical computation. We demonstrate the consistency of these estimators and validate their effectiveness through simulation studies with synthetic data.

S004: Advanced Statistical and Machine Learning Methods in Genetics and Genomics

A Scalable Approach for Dynamic Single-Cell eQTL Analysis

Lin Hou

Tsinghua University

Single-cell eQTL studies have revealed that genetic regulation of gene expression is cell-type specific and dynamically dependent on the cell state. We have developed a statistical framework that maps static and dynamic single-cell eQTLs using a functional mixed model with retrospective association test. Simulations showed that our method controls for type I error rates under

various gene expression models and is powerful in diverse dynamic patterns. Applied to the OneK1K dataset, our method identified 30% more eGenes that are cell-type specific compared to the pseudo-bulk analysis and revealed dynamic eQTL effects in a rare subset of dendritic cells.

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

◆ *Yaowu Liu, Tianying Wang*

Southwestern University of Finance and Economics, Colorado State University

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.

Rare Coding and Noncoding Variants Map 1,342 Diseases and Biomarkers in 490,549 Whole Genomes

Zilin Li

Northeast Normal University

Rare genetic variants are increasingly recognized as important contributors to human trait architecture, with noncoding variants accounting for a substantial portion of the heritability. These variants tend to be less polygenic and more biologically specific than common variants, remaining understudied across large biobanks. Here we analyzed whole genome sequencing (WGS) data from up to 490,549 UK Biobank participants to assess the effects of rare coding and noncoding variants across 1,342 phenotypes, including 944 diseases, 76 clinical biomarkers, and 322 metabolomics traits. We developed and applied STAARpipelinePheWAS, a scalable framework for WGS phenome-wide rare variant association analysis, identifying 49,121 genome-wide significant gene-trait pairs. Our study presents a comprehensive map of noncoding rare variant associations in both disease and biomarker domains. Many associations were undetected in prior exome- or array-based studies and were enriched in drug targets and biologically coherent pathways. All results are publicly accessible through an interactive portal (https://www.staarphewas.org/), offering a foundational resource for rare variant discovery, functional interpretation, and translational genomics.

CausalGRN Maps Causal Gene Regulatory Networks and Predicts Unseen Perturbation Effects from Single-Cell CRISPR Screens

◆ *Bo Yu, Dingyu Liu, Guanghao Qi, Danwei Huangfu, Li Hsu, Ali Shojaie, Wei Sun*

Fred Hutchinson Cancer Center, Sloan Kettering Institute, University of Washington, Sloan Kettering Institute, Fred Hutchinson Cancer Center, University of Washington, Fred Hutchinson Cancer Center

Single-cell CRISPR screens with single-cell RNA sequencing (scRNA-seq) readouts provide genome-scale interventions for mapping gene regulation, yet current analyses either construct association networks without leveraging interventions to infer causality or train black-box predictors that do not consistently outperform strong simple baselines on held-out perturbations. We present CausalGRN, a scalable framework that learns a causal gene regulatory network with oriented regulatory influences from thousands of perturbations and predicts transcriptome-wide responses to unseen perturbations. CausalGRN applies a sparsity-aware correction to estimate a robust network skeleton from scRNA-seq data, orients edges using observed perturbation outcomes, and captures shared perturbation responses with explicit gene-program variables. For prediction, it propagates perturbation signals through the inferred network under graph constraints, yielding path-level explanations of downstream changes. Across ground-truth simulations and multiple large single-cell perturbation datasets, CausalGRN improves causal structure recovery and increases accuracy of unseen perturbation-effect prediction.

S008: Advances in Bayesian Modeling for Big and Complex Data

A Semi-Reference Bayesian Approach to Cell-Type Annotation and Marker Discovery in Single-Cell RNA-seq

◆ *An-Shun Tai, Ying-Ting Huang*

Tsing Hua University, Taiwan, Cheng Kung University

Accurate cell-type annotation and marker gene discovery remain

fundamental challenges in single-cell RNA-sequencing (scRNA-seq) analysis due to high-dimensional data, biological noise, and the frequent absence of reliable reference datasets. We propose a semi-reference Bayesian hierarchical approach that models observed gene counts using a negative binomial distribution to account for overdispersion. A key contribution of this framework is its ability to investigate and identify multiple unknown cell types unique to the query dataset while simultaneously classifying known populations. The model integrates a classification submodel for guided annotation and a clustering submodel for unsupervised inference. Central to this approach is a spike-and-slab prior structure that enables simultaneous marker gene selection by distinguishing cell-type-specific effects from baseline expression. To determine the optimal number of cell clusters, the framework employs the deviance information criterion and Watanabe-Akaike information criterion. The robustness of the model is demonstrated through comprehensive simulation studies—ranging from small-scale to realistic 1,000-gene settings—evaluating performance against varying signal strengths and overlapping marker structures. Our results confirm the model's efficacy in characterizing complex biological structures by successfully identifying novel cell populations not present in the reference data.

Bayesian Sparse Kronecker product decomposition

◆ *Shao-Hsuan Wang, Hsin-Hsiung 'Bill' Huang*

Graduate Institute of Statistics of Central University, Taiwan,, University of Central Florida

The Sparse Kronecker Product Decomposition (SKPD) for tensor data was introduced by Sanyou Wu and Long Feng (2023). This method represents the first frequentist framework designed for signal region detection in high-resolution, high-order image regression problems. Their work demonstrated the strong performance of SKPD in various applications.

In this presentation, we will introduce a Bayesian version of SKPD, referred to as Bayesian SKPD. From a Bayesian perspective, we apply a three-parameter beta-normal prior family to the parameters of interest. Additionally, we address tensor regression data with mixed-type responses using Polya-Gamma augmentation. This approach allows us to give credible region detection through direct Gibbs sampling. The theoretical results will be presented, and we will demonstrate the effectiveness of Bayesian SKPD using real brain imaging data from the OASIS.

Synthesizing Multiple-Type Recurrent Events with the Terminal Event: A Joint Bayesian Copula-Frailty Modelling Approach

◆ *Menglu Liang, Yifan Yang, Lijun Zhang, Chixiang Chen, Ming Wang*

University of Maryland, College Park, Case Western Reserve University School of Medicine, Case Western Reserve University School of Medicine, University of Maryland, Baltimore, Case Western Reserve University School of Medicine

In clinical and observational studies, multiple types of recurrent events are often encountered, such as heart failure, myocardial infarction, stroke, or other cardiovascular diseases, which are likely to be highly correlated with each other. During follow-up, these recurrent events can be censored by a competing risk event (e.g., death), leading to dependent censoring. This multimorbidity phenomenon constitutes a serious challenge for healthcare systems due to its association with poorer health-related outcomes, more complex clinical management, and increased health service utilization and costs. To address these issues, we propose a cohesive framework for modeling fully synthetic high-dimensional time-to-event data of multiple types. Our approach centers around a joint Bayesian model that simultaneously borrows information across all these event types, enabling inference based on a synthetic version of the data. Specifically, the proposed methodology is based on a joint frailty modeling approach with multivariate random effects coupled with a multivariate copula to account for the heterogeneous risk of failure across individuals and the presence of informative censoring due to a terminal event dynamically. We utilize a Bayesian technique based on the Metropolis-Hastings within Gibbs sampling algorithm for parameter estimation and inference. We demonstrate the capacity of our approach using a prospective cohort study with recurrent stroke, heart failure, and MI patients and document the relative performance of our proposal compared to a univariate copula-frailty model via extensive simulation studies. Our new approach is timely for advancing tools to address increasingly complex needs related to multimorbidity and developing interventions that are most effective and viable for personalized medicine.

S009: Advances in High-Dimensional and Complex Data Analysis

The Fairness Hierarchy: A Viewpoint from Causal Inference

Chengbo Zhang, Zhen Yao, Hao Pang, ◆ *Changcheng Li*

Dalian University of Technology, Dalian University of Technology, Dalian University of Technology, Dalian University of Technology

Fairness in machine learning prediction has attracted growing attention in recent years. In this article, we propose a causal–inference–based framework for fair prediction, defined through

path-specific counterfactual interventions. Instead of imposing fairness via constraints on predictive objectives or model parameters, our approach specifies fairness directly at the level of counterfactual prediction semantics. Given a well-specified causal graph, we construct a predictive distribution for the outcome Y using a structural causal model and generate counterfactual predictions by selectively intervening on causal paths emanating from sensitive attributes. By allowing or blocking the propagation of sensitive information along designated paths, possibly involving multiple sensitive sources, our framework induces a hierarchy of interpretable fairness notions, generalizing standard path-specific causal semantics. Our empirical experiments demonstrate how different fairness levels can be instantiated and compared in practice.

Pursuing Homogeneity and Sparsity in Simultaneous Quantile Regression

Zhen Zeng, ◆ *Weixin Yao*

Nanjing University of Finance and Economics, University of California, Riverside

Quantile regression analyzes the impact of predictors on the conditional distribution of the response by focusing on a collection of conditional quantiles rather than a single conditional mean. The main goal of this work is to Pursuing Homogeneity in Simultaneous Quantile Regression (PHISQ) to get more interpretable and efficient quantile regression estimation. The new method reveals not only the predictors that are associated with the outcome, but also the true sources of homogeneity, the specific predictors that only have homogeneous effects on the response at the quantiles of interest, and the true sources of heterogeneity, the predictors that have heterogeneous effects across quantile levels of interest. Therefore, the new method may eventually result in a model that is considerably more parsimonious and interpretable. In addition, the new method can pool/borrow information across different quantiles to estimate the homogeneous regression parameters, and thus improve the estimation efficiency, especially at the tails. We demonstrate that the penalized PHISQ method exhibits desirable properties under mild regularity conditions. Simulation results and a real data application further validate the effectiveness of PHISQ.

Conditional Expectile-Based Value-at-Risk: Semiparametric Estimation and Inference

Feipeng Zhang

School of Economics and Finance, Xi'an Jiaotong University

Conditional expectile-based value-at-risk (EVaR) has been recommended in financial risk management due to its coherency and elicitability. However, the existing nonparametric conditional EVaR estimation suffers from the curse of dimensionality. To this end, we study semiparametric estimation and inference for conditional EVaR in nonlinear time series models with unspecified error distribution. Specifically, we propose a two-step conditional EVaR estimator that achieves the parametric convergence rate. Furthermore, we derive a functional central limit theorem for conditional EVaR process, and then extend to investigate conditional expectile-based expected shortfall and conditional distribution estimation. A sieve bootstrap procedure is proposed for non-expert practitioners to conduct statistical inference. The asymptotic results and their applications are illustrated by numerical studies and an empirical application to S&P 500 returns.

Posterior Predictive (PoP) Design for Next-generation Phase I Clinical Trials

Yeqing Zhou

Tongji University

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.

S017: Advances in Statistical Learning and AI for Biological and Health Sciences

Single-Cell Gene Regulatory Network Analysis with Applications to Alzheimer's Disease

Danni Liu, Zhongli Jiang, Yan Li, Dabao Zhang, ◆ Min Zhang

University of California, Irvine, University of California, Irvine, University of California, Irvine, University of California, Irvine, University of California, Irvine

Alzheimer's disease (AD) involves widespread disruptions in gene regulation across multiple brain cell types; however, the intracellular causal mechanisms remain largely unknown. In

this study, we employed SIGNET to analyze single-nucleus transcriptomic and genotype data from brain tissue, and constructed cell-type-specific causal gene regulatory networks (GRNs). Our approach leverages the power of genomic variants as instrumental variables to establish robust causal relationships between genes through bootstrap datasets. Results include well-known GRNs, new regulations, cell-type-specific hub genes, and pathways involved in AD pathogenesis. These findings expand resources for further functional studies and offer new insights into the cell-type specific molecular mechanisms underlying Alzheimer's disease.

Accelerating with AI: Reflections from a Computer Vision Project in Grass Seed Analysis

Yanming Di

Oregon State University

In this talk, I will reflect on my personal journey using modern AI tools, particularly large language models (LLMs), to accelerate work in an applied research setting. Using a computer vision project in grass seed analysis as a running example, I will focus on a critical one-month period in late 2025 when our team experienced significant turnover and instability. During this time, AI tools played a key role in helping me maintain momentum, enabling rapid refactoring and development of both training pipelines and client-facing components, often working independently in a highly iterative "vibe coding" style.

I will also discuss lessons emerging from the transition as the team later expanded, including the challenges of moving from solo, AI-assisted development to a collaborative, multi-developer environment.

While the landscape of AI tools is evolving rapidly, one key message is that AI has the potential to turn each of us into a more well-rounded, full-spectrum developer. At the same time, I will leave open the question of how statisticians can continue to differentiate themselves and provide unique value in this changing landscape.

How Much Biology Do LLM-Derived Gene Embeddings Actually Know?

Yufei Cai, Dailin Gan, Hui Zhang, ◆ Jun Li

University of Notre Dame, University of Notre Dame, Northwestern University, University of Notre Dame

LLM-derived gene embeddings from brief NCBI gene descriptions have shown strong performance in biological applications, yet what they encode remains unclear. Using a GSEA-based framework, we show that OpenAI's embeddings recover the vast majority of Hallmark and C2 pathways. Surprisingly, embeddings from gene symbols alone still capture substantial pathway information, revealing prior biological knowledge retained from pretraining. A comparison of eleven smaller models shows that even modest textual context is sufficient for all models to approach comparable coverage. These findings demonstrate that LLM-derived gene embeddings encode far richer biology than their brief inputs would suggest.

S020: Advances in Statistical Methods for Omics and Health Data Analysis

A Computational Pipeline for Studying Multicellular Interactions in 3D Bioprint Systems

Min Yang, ◆ Sha Cao

Oregon Health & Science University, Oregon Health & Science University

A major challenge in analyzing multicellular 3D culture systems is that the cell populations initially introduced into the model may change substantially after co-culture. In response to the surrounding microenvironment, cells may differentiate, reprogram, or progress into new states, making annotation based solely on transcriptomic similarity to the input cell lines unreliable. This challenge becomes particularly important when the objective is not only to identify cell types, but also to trace lineage-related transitions and quantify how different stromal or immune compositions reshape cellular states in engineered tissues.

To address this problem, we developed an integrated computational pipeline that combines SNP-based assignment with expression-based annotation. Genotype information, inferred from RNA-sequencing profiles, is used to anchor each cell to its most likely source population, while transcriptomic profiles and curated marker genes are used to refine cell-type and cell-state labels. This hybrid strategy is designed to remain robust in the presence of substantial transcriptional drift, lineage conversion, and intermediate cellular states.

We applied this pipeline to our in-house human 3D bioprint system for multicellular interaction analysis, in which pancreatic cancer cells are co-cultured with stromal and microenvironmental components, including CAF-related populations, mesenchymal and stellate cells, endothelial cells, and immune cells. For this platform, we generated single-cell RNA-seq data from both the mixed 3D tissues and the individual cell lines used as model inputs, creating a controlled system for studying cell identity, plasticity, and tumor–microenvironment interactions. The resulting framework resolves major cellular compartments, including cancer, fibroblast, endothelial, and immune cells, and further distinguishes fibroblast-related populations such as CAF, MSC, PSC, and HSC states.

The pipeline further enables three major downstream analyses: annotation of cell identities in heterogeneous 3D tissues, inference of differentiation, reprogramming, and progression trajectories, and structured comparison of cellular programs across different stromal and immune combinations. In particular, it supports comparisons of the same stromal subtype across distinct cancer contexts, as well as comparisons of cancer cell states under alternative stromal compositions. Overall, this work provides a computationally robust framework for studying multicellular dynamics in bioprint and organoid systems.

Advancing Computational Resources and Tools to Study Metabolism in Diseases

◆ *Chi Zhang, Sha Cao, Pengtao Dang, Paveethran Swaminathan*

Oregon Health & Science University, Oregon Health & Science University, Oregon Health & Science University, Oregon Health & Science University

Metabolic dysregulation is a unifying hallmark of diverse human diseases, including cancer, neurodegeneration, inflammatory disorders, and cardiometabolic syndromes. However, translating high-dimensional omics measurements into quantitative, mechanistic insight about metabolic function remains a major computational bottleneck. Current approaches largely rely on static pathway enrichment or bulk averages, which fail to resolve cell-type–specific, spatially localized, and tissue-to-tissue metabolic interactions that ultimately drive disease phenotypes. To address this gap, we are developing a next-generation computational ecosystem that integrates scalable algorithms, curated metabolic knowledge bases, and physics-informed machine learning models to enable quantitative inference of metabolic states directly from multi-omics data. Our framework reconstructs cell- and tissue-specific metabolic networks, estimates reaction-level flux distributions from transcriptomic and proteomic measurements, and models intercellular and inter-organ metabolite exchange using spatial and systems-level constraints. These tools are designed to operate across modalities—including bulk, single-cell, spatial transcriptomics, metabolomics, and longitudinal clinical data—and to scale to population-level datasets comprising millions of cells. By combining mechanistic modeling with modern AI and cloud-based infrastructure, this platform enables hypothesis generation, in silico perturbation, and digital-twin simulation of disease metabolism. The resulting computational resources will provide the research community with reproducible, interoperable, and extensible tools to systematically characterize metabolic reprogramming, identify therapeutic vulnerabilities, and prioritize biomarkers across disease contexts. Together, this work establishes a generalizable foundation for data-driven systems metabolism and accelerates the translation of multi-omics data into actionable biological and clinical insight.

Correcting Bias from Covariate-Dependent Censoring in Survival Disparity Decomposition

◆ *Yan Li, Chong Zhang, Menglu Liang*

University of Maryland, University of Maryland, University of Maryland

Research on health disparities (HD) in survival outcomes is crucial to public health, yet existing HD decomposition methods can yield biased estimates when censoring depends on covariates—a common feature of observational studies in which loss to follow-up varies with demographic, socioeconomic, or behavioral characteristics. Widely used Peters–Belson (PB)–type approaches exacerbate this issue by relying on the Kaplan–Meier (KM) estimator to quantify overall disparities and on the Cox model to predict counterfactual survival, even though KM assumes unconditional independent censoring while Cox assumes conditional independence, leading to bias under covariate-dependent censoring. To address this, we propose a novel decomposition method that integrates propensity score (PS) methodology with inverse probability of censoring weights (IPCW). The PS component balances covariates across groups to estimate counterfactual survival without specifying an outcome model, while the IPCW component corrects for informative censoring by upweighting uncensored individuals to represent those censored. Developed within a semiparametric framework that models censoring but estimates survival nonparametrically, the proposed IPCW-PS approach corrects biases from traditional methods that rely on KM and Cox. We evaluate performance through simulation studies and apply the method to NielsenIQ Consumer Panel data to quantify and decompose dynamic disparities in smoking cessation between White and Black households.

Inferring Non-Coding RNA Regulatory Network from Transcriptomic Data and Curated Databases

Hongjie Ke

Capital University of Economics and Business

Non-coding RNAs (ncRNAs) have played an indispensable role in regulating gene expression and cellular processes. However, the complex regulatory circuits between ncRNAs and coding genes remain understudied. Major challenges for inferring ncRNA regulatory network (NRN) on a transcriptome-wide scale include high dimensionality of both ncRNA and coding gene expression data, their context-dependent interaction, and lack of validation. We hereby propose a comprehensive analytical framework, namely Construction and Analysis of non-coding RNA regulatory NETwork (CAR-NET), specifically designed to infer NRN from transcriptomic data and curated databases on experimentally validated ncRNA-gene interactions and disease-related ncRNAs. At its core is a novel computationally efficient Bayesian network

structure learning method that leverages the semi-bipartite graph structure of NRN for dimension reduction, edge orientation and global network search. We also developed and applied algorithms to perform differential network, subnetwork and pathway analysis of the inferred network. We benchmarked CAR-NET and showed its strength over existing methods using both synthetic data from simulation and DREAM challenge. We also applied CAR-NET to three case studies, encompassing lncRNA regulation in brain development, miRNA differential regulation in renal cell carcinoma and regulation by different classes of ncRNAs in different cell lines using bulk or single cell RNA-seq data. An R-shiny app was developed to implement the method(s) with interactive GUI.

S027: Applications in Biomedical Sciences

Rank-Based Measures of Liquid Association

Hao Ho

University of California, Los Angeles

Multi-omics data offers unprecedented opportunities to uncover complex associations between molecular features at different levels, many of which are expected to exist only in specific cell types and cellular states. Liquid association (LA) is a valuable tool for integrating multi-omics data, conceptualizing the intrinsic evolution of the association between a pair of variables (X, Y) as a third variable Z varies. However, the normality conditions of the LA pair (X, Y) and Z limit the applicability of LA due to the multi-modality of data types. In this study, we exploit Kendall's tau to redefine LA, allowing (X, Y) to be non-normal continuous, ordinal, or interval variables. We derive two versions of rank-based LA: one global and one conditional, and discuss their relative merits. Additionally, we present nonparametric methods to accommodate the unspecified distribution of Z. Our method is applied to identify triplets with significant liquid association in a heterogeneous multi-omics lung adenocarcinoma dataset from The Cancer Genome Atlas. We anticipate that our method can be applied to other similarly complex datasets across various scientific fields.

CRISP: A Modular Platform for Cryo-EM Image Segmentation and Processing

Szu-Chi Chung

Sun Yat-Sen University, Taiwan

Separating true particle signal from background in cryo-EM micrographs remains a formidable challenge, owing to the characteristically poor signal-to-noise ratio (SNR) of the data, the co-occurrence of contaminants, spatially varying ice thickness, and the presence of heterogeneous, densely arranged particles. Pixel-level segmentation approaches have emerged as a compelling alternative to classical detection-based methods, offering superior false-positive suppression through blob-mass filtering, more precise particle centering via full intensity-profile exploitation, and enhanced sensitivity to morphologically irregular targets. Nevertheless, the acquisition of reliable pixel-level annotations under low-SNR conditions remains a fundamental obstacle, and the absence of a unified benchmarking framework has left the field fragmented, with most existing pipelines built around task-specific, ad hoc design decisions.

To address these limitations, we introduce a modular, end-to-end platform for the automated generation of high-fidelity reference segmentation maps. The framework accommodates flexible combinations of segmentation backbones, feature extraction modules, and loss objectives, and incorporates a novel Conditional Random Fields (CRF) refinement module that leverages class-discriminative representations within a regularized optimization scheme to transform coarse probability maps into precise, fine-grained segmentation masks. We further develop a post-processing pipeline that converts predicted segmentation masks into particle coordinates, bridging pixel-level predictions with downstream picking workflows. On synthetic benchmarks, models supervised with our reference annotations consistently exceed 90% across all major pixel-level metrics — accuracy, precision, recall, IoU, and F1 score. Crucially, we validate the utility of the resulting segmentation maps on real cryo-EM data, demonstrating that particle sets derived from our predictions yield three-dimensional density reconstructions of quality comparable to expert-curated picks and surpassing those obtained by leading automated particle-picking tools.

Inference for Heavy-Tailed Integer-Valued Time Series with Applications to Infectious Disease Case Counts

Yuichi Goto, ◆ *Gaspard Bernard*

Kyushu University, Academia Sinica, Taiwan

Integer-valued time series models arise naturally in the analysis of infectious disease case counts. It is well known that such data often exhibit specific features, including zero inflation and over-dispersion. For INMA(1) and INAR(1) models, we propose a parametric approach based on the generalized spectrum of the process. The proposed estimators are robust to the lack of finite moments of order 1. We study their asymptotic properties, establishing root-n consistency and asymptotic normality under discrete-stable innovations, which allows us to accommodate both zero inflation and heavy tails. We illustrate the practical relevance of the method through applications to parameter estimation and forecasting for weekly case counts of several infectious diseases in Germany, namely Measles and the E. coli and EHEC bacteria. The datasets considered display varying degrees of over-dispersion and zero inflation. Overall, the proposed approach provides a

flexible framework for estimation, goodness-of-fit assessment, and forecasting for these three markedly different diseases.

Toward Robust Machine Learning Methods for Individualized Treatment Effect Estimation

Keng-Te Liao

Tsing-Hua University, Taiwan

Reliable estimation of treatment effects is fundamental to advancing precision medicine and clinical decision-making. While machine learning (ML) approaches are favored for their ability to model complex and high-dimensional data, they often suffer from sensitivity to data noise, leading to inconsistent or non-robust estimations. In this talk, I will present a novel framework that identifies both the algorithmic and theoretical bottlenecks inherent in current estimation methods. I will discuss how this framework addresses these robustness challenges and provides a blueprint for designing more effective model training paradigms. Finally, I will demonstrate a streamlined training approach derived from this framework that significantly improves estimation performance across diverse scenarios.

S046: Frontiers in Causal Inference: Addressing Time-Dependence, Fairness, and Long-Term Effects in Complex Data

Ensuring Fairness in Sequential Decision Making: Counterfactual Fairness in Reinforcement Learning for Opioid Use Disorder Services

Zhenke Wu

University of Michigan

As healthcare systems move to deploy reinforcement learning (RL) for dynamic resource allocation, significant concerns have been raised about the impact of these tools on health services access and equity. While RL may improve service personalization, black-box optimization based on historical data can inadvertently lead to withholding needed treatment from vulnerable subgroups. We investigated the impact on fairness in treatment allocation using clinical data from a study that tested an RL-supported intervention for managing chronic pain and opioid misuse risk. In this talk, I show how to (1) audit standard RL policies for allocation bias across patient subgroups, and (2) validate a deployable methodology for correcting identified disparities without compromising clinical outcomes. We show that standard methodologies seeking to achieve "fairness through unawareness" are insufficient. Our framework evaluates and patches these vulnerabilities prior to deployment, satisfying a key requirement for the ethical scaling of AI in healthcare. Although motivated by health applications, the proposed framework applies in broader operations management contexts.

On robust covariate adjustment methods in RCT

Lin Liu

Shanghai Jiao Tong University

Covariate adjustments have been important in improving the efficiency of treatment effect estimates in the context of RCT, even though the naive unadjusted estimator is $\sqrt{n}$-consistent and asymptotic normal with variance consistently estimable from data. Covariate adjustments have been recommended in FDA guidelines, when the number of basline covariates is sufficiently small compared to the sample size. However, when the number of baseline covariates is relatively large ($p \gtrsim \sqrt{n}$), the recommendation is murky due to bias inflation. In this talk, we demonstrate that a simple jackknife-based covariate adjustment procedure that can achieve efficiency gain without deterioriating the bias, as long as $p = o (n)$. This procedure not only works for linear working models, but also for a very large class of nonlinear working models, including GLMs, Single-Index Models, boosting, and even neural networks. In particular, the unbiasedness holds for any p and even when $p = \infty$. Our unified procedure is the first that does not rely on sample splitting in the spirit of double machine learning (Chernozhukov et al. 2018), which is not recommended for RCT analysis due to the dependence of the final results on how the split is done.

Causal Mediation Analysis in Multistate Process

Yiqing Chen

Texas A&M University

Unraveling the complex relationships among interventions, intermediate health states, and survival outcomes is crucial to translating advances in clinical research into optimal therapies. We propose a novel Bayesian multi-state mediation model to evaluate the direct and indirect effects of treatment modalities on overall survival (OS). Our model extends traditional mediation analysis by incorporating a multistate framework that accounts for intermediate outcomes, such as drug response and adverse events, providing a more detailed representation of treatment impact over time. Bayesian estimation and variable selection enhance model reliability and address uncertainty in parameters and structure, providing robust estimates of direct and mediated effects. The method characterizes dynamic relationships between endpoints and intermediate health states which may be time-varying mediators. Our study highlights the clinical importance of considering intermediate outcomes in the evaluation of treatment efficacy and patient survival. The model is applied to a propensity-matched retrospective cohort. Results reveal the

observed survival benefit is significantly attributable to reduced risk of adverse event, and provide a dynamic prediction of survival probability by time.

A General Framework for Incorporating Identification Uncertainty in Individualized Treatment Rules

Muxuan Liang, Ting Ye, ◆ Yingqi Zhao

MD Anderson Cancer Center, University of Washington, Fred Hutchinson Cancer Center

Estimating individualized treatment rules (ITRs) from observational data or clinical trials with non-adherence is challenging due to potential unmeasured confounding bias. An instrumental variable (IV) can be used to partially identify possible values for conditional average treatment effects (CATEs).

However, optimal treatment decisions under partially identified CATEs can be uncertain; and current literature fails to inform such uncertainty in treatment decisions. We propose a novel framework incorporating a reject option to develop new classes of 'optimal' ITRs that guide treatment decisions while accounting for identification uncertainty. This reject option highlights patients susceptible to identification uncertainty and facilitates the collection of additional information to aid in optimal decisions. We introduce a unique loss function linking the reject option with uncertain treatment decisions and derive associated IV-optimal ITRs. Our framework allows users to control the size of subgroups receiving the reject option by considering the cost of collecting additional information versus the risk of incorrect treatment decisions. We utilize a weighted classification framework with a modified hinge loss function and propose an augmented empirical risk minimization approach to estimate the IV-optimal ITRs, achieving fast convergence even with nonparametric or machine learning methods for estimating nuisance parameters.

S053: High-Dimensional Statistics

A Non-Universal Convex Non-Gaussian Min–Max Theorem and Its Implications for High-Dimensional Machine Learning

Mohamed Chiheb Yaakoubi, Cosme Louart, Malik Tiomoko, ◆ Zhenyu Liao

Chinese University of Hong Kong, Shenzhen, Chinese University of Hong Kong, Shenzhen, Huawei, Huazhong University of Science and Technology

We study high-dimensional convex empirical risk minimization (ERM) under general non-Gaussian data designs. By heuristically extending the Convex Gaussian Min–Max Theorem (CGMT) to non-Gaussian settings, we derive an asymptotic min–max characterization of key statistics, enabling approximation of the mean and covariance of the ERM estimator. Specifically, under a concentration assumption on the data matrix and standard regularity conditions on the loss and regularizer, we show that for a test covariate independent of the training data, its projection on the ERM estimator approximately follows the convolution of a (generally non-Gaussian) distribution with an independent centered Gaussian variable. This result clarifies the scope and limits of Gaussian universality for ERMs.

Tensor Elliptical Graphical Model

Jixuan Liu, Zhengke Lu, ◆ Le Zhou, Long Feng, Zhaojun Wang

Nankai University, Nankai University, Hongkong Baptist University, Nankai University, Nankai University

We address the problem of robust estimation of sparse high dimensional tensor elliptical graphical model. Most of the research focus on tensor graphical model under normality. To extend the tensor graphical model to more heavy-tailed scenarios, motivated by the fact that up to a constant, the spatial-sign covariance matrix can approximate the true covariance matrix when the dimension turns to infinity under tensor elliptical distribution, we propose a spatial-sign-based estimator to robustly estimate tensor elliptical graphical model, the rate of which matches the existing rate under normality for a wider family of distribution, i.e. elliptical distribution. We also conducted extensive simulations and real data applications to illustrate the practical utility of the proposed methods, especially under heavy-tailed distribution.

Adaptive Detection of Change-Points for High-Dimensional Covariance Matrices

◆ *Jiayu Lai, Xiaoyi Wang, Le Zhou, Shurong Zheng*

Jiangxi University of Science and Technology, Beijing Normal University, Hong Kong Baptist University, Northeast Normal University

This paper develops a series of U-statistic-based tests for detecting change-points in high-dimensional covariance matrices under the assumption of temporal independence. The asymptotic distributions of the proposed U-statistics are derived under both the null and local alternative hypotheses. A family of maximum-type statistics is then introduced, followed by two testing procedures based on combining the p-values of these maximum-type statistics. We further propose methods for estimating the change-point location and establish their corresponding convergence rates. Leveraging these estimators, we construct an adaptive detection method for single change-point. An improved version of the detection procedure is subsequently developed to identify multiple change-points by using the binary segmentation technique. Simulation studies demonstrate that the proposed tests maintain high power under alternatives with varying levels of sparsity, and that our adaptive estimator perform well under both

single and multiple change-point scenarios.

Spectral Properties of High-dimensional Rescaled Sample Correlation Matrices under Elliptical Distributions

Weijiang Chen, Shurong Zheng, ◆ *Tingting Zou*

Jilin University, Northeast Normal Univeristy, Jilin University

Under the convergence regime that the data dimension and sample size tend to infinity proportionally, we establish the joint central limit theorem for the linear spectral statistics of the rescaled sample correlation matrices when the population follows an elliptical distribution. This general theoretical result has wide applications. As an illustration, we consider testing whether a high-dimensional correlation matrix equals to a prespecified matrix. We first propose three test statistics based on the quadratic norm and then derive their limiting distributions. To enable the practical applications of the proposed tests, we also develop a numerical approach for computing the asymptotic variances and covariances in the central limit theorem. Furthermore, extensive simulation studies are systematically conducted to evaluate the finite sample performance of the proposed testing procedures.

S067: Modern Mixture Model Methodology: Advances in Modeling, Inference, and Applications

Gaussian Mixture Reduction Approach to Compacting 3D Gaussian Splatting in Computer Graphics

Tao Wang, Mengyu Li, Geduo Zeng, Cheng Meng, ◆ *Qiong Zhang*

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

3D Gaussian Splatting (3DGS) has become a leading technique for neural rendering, representing complex 3D scenes as a mixture of millions of anisotropic Gaussians. While highly effective, the model is inherently over-parameterized, leading to significant computational and memory burdens. Current compaction methods are largely heuristic, relying on local pruning rules that lack a global fidelity guarantee. In this work, we reframe the problem as one of Gaussian Mixture Reduction (GMR), providing a principled statistical framework for approximation. Our GMR-based algorithm efficiently reduces the number of components by orders of magnitude while minimizing the loss of geometric and photometric information. On benchmark datasets, our method maintains rendering quality comparable to the original 3DGS approach using only 10% of the components, and consistently outperforms state-of-the-art compaction techniques. A key advantage of our approach is its agnosticism to the underlying 3DGS pipeline, offering a versatile, efficient, and statistically-grounded pathway to lightweight neural rendering.

Model Positive and Unlabeled Data with a Generalized Additive Density Ratio Model

◆ *Peijun Sang, Yifan Sun, Qinglong Tian, Donglin Zeng, Pengfei Li*

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

This paper focuses on learning from positive and unlabeled (PU) data, where only some positives are labeled and the rest are mixed with negatives. Classical exponential tilting models guarantee identifiability by imposing a linear structure, but they can be severely misspecified when the true relationships are nonlinear. We propose a generalized additive density-ratio framework that retains identifiability while allowing nonlinear and feature-specific effects. The approach comes with a practical fitting algorithm and supporting theory that enable estimation and inference for the mixture proportion and other quantities of interest. In simulations and analyses of real datasets, the proposed method matches the standard exponential tilting method when the linear model is correct and delivers clear gains when it is not. Overall, the framework strikes a useful balance between flexibility and interpretability for PU data and provides principled tools for estimation, prediction, and uncertainty assessment.

Smoothed Likelihood for Combining Diagnostic Tests

Fangyong Zheng, Pengfei Li, ◆ *Tao Yu*

National University of Singapore

In medical diagnostics, leveraging multiple biomarkers can significantly improve classification accuracy compared to using a single biomarker. While existing methods based on exponential tilting or density ratio models have shown promise, their assumptions may be overly restrictive in practice. In this paper, we adopt a flexible semiparametric model that relates the density ratio of diseased to healthy subjects through an unknown monotone transformation of a linear combination of biomarkers. To enhance estimation efficiency, we propose a smoothed likelihood framework that exploits the smoothness in the underlying densities and transformation function. Building on the maximum smoothed likelihood methodology, we construct estimators for the model parameters and the associated probability density functions. We develop an effective computational algorithm for implementation, derive asymptotic properties of the proposed estimators, and establish procedures for estimating the receiver operating characteristic (ROC) curve and the area under the curve (AUC). Through simulation studies and a real-data application, we demonstrate that the proposed method yields more accurate and efficient estimates than existing approaches.

Likelihood-Based Non-Parametric Receiver Operating Char-

acteristic Curve Analysis in the Presence of Imperfect Reference Standard

Yifan Sun, Peijun Sang, Qinglong Tian, ◆ Pengfei Li

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

In diagnostic studies, researchers frequently encounter imperfect reference standards with some misclassified labels. Treating these as gold standards can bias receiver operating characteristic (ROC) curve analysis. To address this issue, we propose a novel likelihood-based method under a non-parametric density ratio model. This approach enables the reliable estimation of the ROC curve, area under the curve (AUC), partial AUC, and Youden's index with favorable statistical properties. To implement the method, we develop an efficient expectation-maximization (EM) algorithm. Extensive simulations evaluate its finite-sample performance, showing smaller mean squared errors in estimating the ROC curve, partial AUC, and Youden's index compared to existing methods. We apply the proposed approach to a malaria study.

S080: New Advances in Microbiome Data Analysis and Related Fields

CAFT: A Compositional Log-Linear Model for Microbiome Data with Zero Cells

Mo Li, Glen Satten, ◆ Ni Zhao

University of Louisiana at Lafayette, Emory University, Johns Hopkins University

Background Differential abundance analysis is fundamental to microbiome research and provides valuable insights into host-microbe interactions. However, microbiome data are compositional, highly sparse (with many zero counts), and influenced by differential experimental biases across taxa. Standard statistical methods often overlook these features. Many approaches analyze relative abundances without accounting for compositionality or rely on pseudocounts, potentially leading to spurious associations and inadequate false discovery rate (FDR) control.

Methods We introduce a novel framework for differential abundance analysis of microbiome data: the Compositional Accelerated Failure Time (CAFT) model. CAFT addresses zero read counts by treating them as censored observations that are below a detection limit. This approach is inherently resistant to multiplicative technical bias, eliminates the need for pseudocounts, and addresses compositional bias through the establishment of appropriate score test procedures.

Results Extensive simulations show that CAFT outperforms competing compositional differential abundance methods, including LOCOM, LinDA, ANCOM-BC2, its robust variant, and LDM-clr by offering more robust type I error and FDR control with or without technical bias. Additionally, we applied CAFT to microbiome data on inflammatory bowel disease (IBD) and the upper respiratory tract (URT) to identify differentially abundant gut microbial taxa between IBD patients and healthy controls, as well as URT taxa distinguishing smokers from non-smokers.

Conclusion We present CAFT, a powerful, robust, and efficient approach for compositional differential abundance analysis. CAFT effectively controls Type I error and maintains FDR control, while demonstrating enhanced power in statistical testing. These capabilities render CAFT a useful tool for compositional microbiome data analysis.

Mediation Analysis with Compositional Exposures

Jing Ma

Fred Hutchinson cancer center

Understanding the causal pathways linking the gut microbiome to downstream biomarkers and clinical outcomes is central to microbiome research. When the microbiome acts as the exposure, the high dimensionality and compositional structure of the data introduce unique methodological challenges for mediation analysis. To address these challenges, we propose a latent variable mediation framework that captures variation in microbial composition through a microbial balance, defined as the log-ratio between two unknown subsets of taxa. This balance serves as a latent scalar exposure and simplifies the estimation of the overall indirect effect at the community level, while simultaneously identifying specific taxa that contribute to the overall direct and indirect effects. This talk will cover the model's estimation and inference, and we illustrate the method using data from the Multiethnic Cohort Study, where we examine the mediation pathway from the gut microbiome through lipopolysaccharide-binding protein to C-reactive protein, an inflammation marker.

Joint Representation Learning, Feature Selection, and Clustering for Sparse Count Data with Applications to Single-Cell Genomics

Cheng Zhong,Siqi Jiang, ◆ Zhi Wei

New Jersey Institute of Technology,New Jersey Institute of Technology,New Jersey Institute of Technology

High-dimensional, sparse, and noisy data are common in many scientific statistical learning applications, where clustering is often used to discover latent structure. A key challenge is that feature selection and clustering are typically treated as separate steps: features are selected first, and clustering is performed afterward. This decoupled strategy can be suboptimal because the selected features are not necessarily the most informative for the downstream clustering objective. In this talk, I will present FSSC, a unified learning framework that performs feature selection

and clustering jointly. Although motivated by single-cell RNA sequencing data, the core idea is methodological: FSSC integrates representation learning, structured sparsity, and clustering supervision into a single optimization framework. Specifically, the model uses a probabilistic autoencoder to capture the distributional properties of sparse count data, incorporates a group Lasso penalty to select a compact subset of informative features, and introduces a clustering-oriented loss to preserve latent cluster structure. By optimizing these components together, FSSC learns low-dimensional representations while simultaneously identifying features that are directly relevant to cluster separation. Using single-cell RNA sequencing as an application domain, we show that FSSC improves clustering accuracy over state-of-the-art methods and identifies compact, interpretable feature sets corresponding to biologically meaningful marker genes. More broadly, this work illustrates how task-aware feature selection can be integrated with deep representation learning to improve unsupervised discovery in high-dimensional scientific data.

Estimating equations with inverse probability weighting for microbiome analysis

Zhigang Li

University of Florida

Zero-inflation is a common problem in microbiome studies where typically around 50\% of the microbiome sequencing reads could be zero. Many existing methods require imputing zeros by a positive number such as Pseudocount to make inference. Imputation approaches are problematic because there is no consensus on what values should be used for replacing zeros. In this paper, we propose a novel approach to address the issue by employing inverse probability weighting so that data points with higher probability of being observed as zeros have more weights in the analysis. It is similar to treating zeros as missing data. This new approach can also handle the compositional structure of microbiome data and make inference on the absolute abundance of an ecosystem. An extensive simulation study is conducted to demonstrate the performance of the proposed method in comparison with existing approaches. A real study application is also presented to showcase the comparison with existing methods in real-world settings.

S089: Random Matrix Theory and Applications

Tuning-Free Multiplier Bootstrap for Edge Eigenvalues of Large Covariance Matrices

◆ *Xiucai Ding, Yichen Hu, Jiahui Xie*

University of California, Davis, University of California, Davis, University of California, Davis

In this talk, I will introduce a novel statistical inference procedure for the edge eigenvalues of large sample covariance matrices based on the multiplier bootstrap. The proposed bootstrap approach is built upon a carefully designed class of feasible multipliers, which allows the method to operate in a fully tuning-free manner. The procedure is particularly well suited for inference at the spectral edge, where classical methods often require delicate calibration or rely on strong assumptions. By avoiding any tuning parameters, the proposed method enhances both practical applicability and theoretical robustness. As an application, the method can be used to construct confidence intervals for non-spiked edge eigenvalues. These confidence intervals, in turn, provide a principled and data-driven approach to estimating the number of spikes in high-dimensional covariance structures, without the need for additional thresholding or tuning procedures.

Spectral Analysis of High-Dimensional Fisher Matrices Built from Sample Correlations

◆ *Xiaoyu Zhang, Shurong Zheng, Xue Ding*

Northeast Normal University, Northeast Normal University, Jilin University

In many high-dimensional problems, variables may have very different variances, making correlation-based normalization more appropriate than covariance-based methods. This leads naturally to Fisher matrices constructed from two independent sample correlation matrices. However, the nonlinear self-normalization in correlation matrices creates substantial technical challenges, and their asymptotic spectral behavior is still not well understood.

In this talk, I will discuss the spectral behavior of such Fisher matrices under a spiked population correlation model. I will show that their empirical spectral distribution converges to the classical Wachter law, so the global limiting spectrum remains the same as in the covariance-based setting. I will then describe the phase-transition phenomenon for extreme eigenvalues, including the precise threshold for spike separation and the deterministic limits of the outliers. In addition, I will present a central limit theorem for the outlier eigenvalues, showing that, after suitable centering and scaling, their joint fluctuations are governed by the eigenvalues of a Gaussian matrix . The proofs rely on a Taylor expansion of the self-normalization terms, precise asymptotic analysis of block bilinear forms, and a martingale-difference CLT.

Extreme eigenvalue gaps in random matrices

Renjie Feng, Jiaming Li, ◆ *Dong Yao*

The University of Sydney, Chinese Academy of Sciences, Jiangsu Normal university

In the talk, I will give a brief review of the extreme gap problems (smallest and largest gaps of the eigenvalues) of various random

matrix ensembles. I will also present our recent series of work on smallest gaps of several random point processes associated with Gaussian structures.

S006: Advanced Statistical Methods for Network Data and Latent Variable Models

Network Goodness-of-Fit for the Block-Model Family

Jiashun Jin, Tracy Ke, Jiajun Tang, ◆ *Jingming Wang*

Carnegie Mellon University, Harvard University, Harvard University, University of Virginia

The block model family is widely used in network modeling and includes four popular models: SBM, DCBM, MMSBM and DCMM. However, the question of which block model best fits real networks has received limited attention in the literature. In this talk, I will introduce a novel approach using cycle count statistics to address the Goodness-of-Fit for these block models. By leveraging the cycle count statistics and a network fitting scheme, we construct four GoF metrics with parameter-free limiting distributions of N(0,1) under the assumed models. We apply these GoF-metrics to some frequently-used real networks for comparison. The numerical results suggest that DCMM is particularly promising for modeling undirected networks.

Community Detection in Heterogeneous Signed Networks

Yuwen Wang, Shiwen Ye, □ *Jingnan Zhang, Junhui Wang*

The Chinese University of Hong Kong, The Chinese University of Hong Kong, University of Science and Technology of China, The Chinese University of Hong Kong

Network data has attracted growing interest across scientific domains, prompting the development of various network models. Existing network analysis methods mainly focus on unsigned networks, whereas signed networks, consisting of both positive and negative edges, have been frequently encountered in practice but much less investigated. In this paper, we formally define strong and weak balance in signed networks, and propose a signed block β-model, which is capable of modeling strong- and weak-balanced signed networks simultaneously. We establish the identifiability of the proposed model by leveraging properties of bipartite graphs, and develop an efficient alternating updating algorithm to optimize the resulting log-likelihood function. More importantly, we establish the asymptotic consistencies of the proposed model in terms of both probability estimation and community detection. Its advantages are also demonstrated through extensive numerical experiments and the application to a real-world international relationship network.

Inference on Generalized Latent Variable Models with High-Dimensional Responses and Covariates

◆ *Jing Ouyang, Chengyu Cui, Yunxiao Chen, Kean Ming Tan, Gongjun Xu*

University of Hong Kong, University of Michigan, London School of Economics and Political Science, University of Michigan, University of Michigan

Regression models with both high-dimensional responses and covariates have attracted growing attention. Standard multivariate regression models become inadequate when the response variables depend not only on observed covariates but also on latent variables that capture key unobserved characteristics. To draw statistical inferences on covariate effects while accounting for latent variables, we consider a high-dimensional generalized latent variable model that accommodates mixed-type responses and allows for flexible dependence between covariates and latent variables, which is more suitable for many real-world applications than traditional methods that either rely on a linear regression form or restricted assumptions on the dependence between covariates and latent variables. We develop an alternating algorithm that iteratively updates the regression parameters and the latent variables, transforming an intractable nonconvex problem into a sequence of tractable convex subproblems. Theoretically, we provide algorithmic guarantees by deriving error bounds for the resulting estimator, proving statistical consistency, and characterizing its convergence rate. Further, building on this estimator, we construct debiased estimates for the covariate effects and establish their asymptotic normality.

Directed Network Discovery with Latent Variables: Identifiability and Estimation

◆ *Siliang Zhang, Yunxiao Chen, Irini Moustaki*

East China Normal University, London School of Economics and Political Science, London School of Economics and Political Science

Learning causal network structures from observational data, especially with latent variables, is challenging. We introduce a framework using general Structural Equation Models (SEMs) for this task, accommodating both latent endogenous and exogenous variables. To ensure the learned latent structure is a Directed Acyclic Graph (DAG), we employ a continuous DAG characterization, enabling smooth, gradient-based optimization. The structure learning is cast as a continuous optimization problem under this DAG constraint, solved efficiently via an augmented Lagrangian method. We discuss conditions for identifying the latent causal structure. This approach offers a flexible, theoretically robust, and computationally efficient method for discovering sparse latent causal networks, with its performance demonstrated through simulation studies and real data applications.

S094: Recent Advances in AI for Biological Sciences

Scaling Spatial Transcriptomics: Multiscale Virtual Modeling of Tissue Microenvironments

Zihao Chen

Peking University

Spatial transcriptomics (ST) has highlighted the critical role of tissue microenvironments in both health and disease, yet its high cost and limited scalability restrict its use in large population studies. In contrast, histopathology is inexpensive and widely accessible, but current histology-based methods for inferring spatial gene expression often struggle to accurately recover quantitative expression levels across samples or to capture complex microenvironmental structures. Here, we introduce a population-scale framework that reconstructs virtual spatial transcriptomic profiles and tissue microenvironments by integrating histology images with bulk RNA sequencing data. By incorporating global molecular signals from bulk RNA, the approach resolves inherent ambiguities in expression magnitude and enables robust quantitative comparisons across cohorts. Applied to data from The Cancer Genome Atlas, the framework generates more than 4,000 virtual spatial transcriptomic profiles at single-cell resolution across multiple cancer types. These data support systematic identification of conserved and prognostic tissue microenvironments, including immune- and stroma-associated niches linked to patient outcomes and therapeutic responses. Overall, this work advances population-scale spatial biology and establishes a foundation for digital tissue models and precision medicine.

Mapping the landscape of histomorphological cancer phenotypes with self-supervised learning

Ke Yuan

University of Glasgow

Self-supervised representation learning is transforming computational pathology, enabling powerful predictive models for cancer diagnosis, prognosis, and treatment response. We leverage this technology to systematically map recurrent and rare histomorphological patterns directly from unannotated pathology slides. Across multiple cancer types, we found previously underappreciated histological features strongly correlated with patient outcomes and molecular data. The patterns define a quantitative landscape of cancer phenotypes, facilitating novel hypothesis generation and deeper biological understanding. This work demonstrates the power of self-supervised AI to unlock clinically relevant insights and discoveries from routinely collected pathology data.

Theory and Algorithms for Constructing AI-Based Dynamic Virtual Cells

Peijie Zhou

Peking University

Artificial Intelligence Virtual Cell (AIVC) is increasingly emerging as a frontier in the interdisciplinary integration of biology and artificial intelligence. Its core vision is to construct digital twins capable of simulating and predicting the dynamic evolution of cellular states, thereby providing computational support for experimental design and mechanistic analysis (Cell 2024). Although omics foundation models inspired by Large Language Models (LLMs) have demonstrated potential in multi-task learning in recent years, recent evaluations indicate that they still face significant limitations in dynamic predictive capability and mechanistic interpretability, particularly when modeling complex biological processes (Nature Methods 2025). In this context, mathematical modeling methods that integrate biological priors and dynamic mechanisms are regaining attention (Cell 2025). However, the key data required for AIVC construction—single-cell multi-omics observations—commonly confront challenges such as sparse sampling and high-dimensional heterogeneity. As a result, traditional modeling approaches centered on differential equations encounter issues of insufficient observability and the curse of dimensionality in data-driven modeling, making it difficult to accurately depict high-resolution, continuous spatiotemporal dynamic processes.

To address these challenges, we have explored a unified modeling framework that integrates generative artificial intelligence methods and dynamical systems theory. By combining mathematical theories such as Optimal Transport, Schrödinger Bridge, and differential geometry with generative AI techniques like Flow Matching and diffusion models, this framework effectively infers continuous dynamic processes of complex state transitions—such as cell proliferation, apoptosis, differentiation, migration, and interactions—from static, heterogeneous single-cell omics temporal snapshots. Compared to black-box methods, this approach not only exhibits strong generative and generalization capabilities but also offers improved mechanistic interpretability. It enables the generation of cell state data across temporal scales and spatial structures, providing a promising direction for constructing dynamic virtual cell models with interpretability, predictive power, and the ability to integrate biological priors.

S095: Recent Advances in Complex Data Analysis

Challenges and Solutions in Comparative Safety Profiles of COVID-19 Vaccines

Tsung-Shan Tsou

Graduate Institute of Statistics of Central University, Taiwan

In the COVID-19 pandemic, various vaccines were manufactured to prevent contracting the disease. Individuals often inject more than one vaccine of different brands for better protection, and alongside each vaccine are multiple side effects.

We introduce a robust score test to determine whether two or more vaccines have the same probability of causing multiple binary side effects.

One can conveniently implement the innovative methodology without dealing with the model encompassing numerous joint probabilities of no direct relevance to the inferences of interest. Our novel technique is applicable when patients have different numbers of side effects and have no paired counterparts.

Piecewise Fréchet Regression for Distributional Data

◆ *Jeng-Min Chiou, Yu-Ting Hsu*

Taiwan University, Taiwan University

We study regression with distribution-valued responses when the relationship may change across a scalar predictor. We propose a piecewise Fréchet regression framework in the Wasserstein space, where conditional means are modeled as Fréchet means and both breakpoints and regime-specific functions are estimated from data.

We outline large-sample properties of the estimators and discuss how breakpoint behavior differs under smooth versus abrupt changes. A cross-validation approach is used for model selection, and a transformation-based alternative is also considered. Simulations and data examples illustrate the method and its practical interpretability.

Stacked Ensemble Prediction for Polynomial Regression with Heteroscedastic Measurement Errors

Jia-Ren Tsai

Fu Jen Catholic University

Polynomial regression with heteroscedastic measurement error is often numerically unstable, and its empirical performance is frequently influenced more by the conditioning of the estimating equations than by sample size alone. In this study, we revisit polynomial errors-in-variables regression from an ensemble perspective. Motivated by methodological grouping, we propose a structured two-layer ensemble framework that first integrates closely related estimators and then constructs a low-variance meta-learner using validation data. Simulation studies under measurement error settings show that unregularized moment-based corrections may exhibit substantial tail risk when the estimating equations are poorly conditioned, whereas ridge stabilization can substantially improve coefficient estimation accuracy. For prediction, the proposed ensemble remains consistently competitive across a wide range of scenarios and typically attains the lowest mean squared prediction error. A real data application is also presented to illustrate the practical value of the proposed method.

Integration of Individual Participant and Aggregate Data Under Dataset Shift

Ming-Yueh Huang

Academia Sinica, Taiwan

Integrated IPD–AD analysis, which combines individual participant data with aggregate data, is widely used for synthesizing evidence across heterogeneous studies. This talk examines how the form of aggregate data affects integration efficiency, an issue that has received less attention than algorithmic development. Using a constrained maximum likelihood framework, I show that subgroup-specific aggregate summaries substantially improve estimation efficiency, with outcome-stratified summaries consistently outperforming covariate-stratified ones, especially for continuous outcomes. Although outcome-stratified summaries are standard for discrete outcomes, they are rarely reported for continuous endpoints; our results suggest that routinely providing such summaries could meaningfully enhance evidence synthesis. I further extend the framework to accommodate dataset shift and propose a fast, non-iterative estimation procedure, illustrated with applications to income data under covariate shift and housing data under prior probability shift.

S101: Recent Advances in Longitudinal Data Analysis

Semiparametric Joint Modeling for Survival Analysis with Longitudinal Covariates

◆ *Wensheng Guo, Tianhao Wang*

University of Pennsylvania, Rush University

Characterizing the association between survival time and the dynamic patterns of a longitudinal covariate trajectory is of particular interest in many studies. Classical time-dependent survival models focus mainly on the link between the concurrent covariate value and the instantaneous hazard function. Consequently, the conditional survival function is often not properly defined on the whole time range, which causes difficulty in model estimation and interpretation. In this article, we propose a novel semiparametric joint modeling approach, in which the observed longitudinal trajectory is modeled as a random realization of a latent functional pattern. We assume each latent pattern uniquely indexes a global survival function via a log-

linear functional regression model. Because the observational time interval of the longitudinal data depends on the survival time, we propose to jointly model the longitudinal and survival data. By using the latent pattern as an infinite-dimensional shared parameter, our approach extends the classical parametric joint modeling method to a semiparametric setting. We show that the proposed estimator achieves the semiparametric efficiency bound. Simulation studies and a real data application demonstrate the advantageous finite sample performances of our new approach.

Varying Treatment Effects in Subgroups: A Unified Framework for Longitudinal Data Analysis

Pu Zhang, Xinsheng Zhang, Jiao Jin, ◆ *Jiakun Jiang*

Beijing Normal University, Beijing Normal University, Beijing Normal University, Beijing Normal University

The rise of precision medicine has shifted statistical analysis from population-averaged treatment effects to subgroup-specific effects, highlighting the importance of tailoring treatments to distinct patient subgroups. Longitudinal data, in particular, provides an opportunity to study how treatment effects evolve over time. However, methods for analyzing subgroup-specific, time-varying treatment effects in longitudinal settings remain limited. Most current approaches rely on two-step procedures with restrictive assumptions, such as a fixed number of subgroups or pre-specified correlation structures. In this paper, we propose a novel semiparametric subgroup-varying-coefficient model that simultaneously estimates all unknowns, including the number of subgroups and subgroup-specific treatment effects, directly from longitudinal data without requiring restrictive assumptions. We develop a penalized Expectation-Maximization (EM) algorithm that simultaneously identifies the subpopulations and estimates the treatment effects in different subgroups. We establish the theoretical properties of the proposed estimators, including consistency and asymptotic normality. The efficacy and robustness of the method are demonstrated through extensive simulations and applications to two real-world datasets: the Standard and New Antiepileptic Drugs study and the Mayo Clinic trial on primary biliary cirrhosis. In both cases, our approach successfully identifies meaningful patient subgroups and reveals subgroup-specific, time-varying treatment sensitivities.

Low-Rank Approximation and Detection for Structurally Changed Tensor Sequence

Jiaqi Huang, ◆ *Ning Wang, Lixing Zhu*

Beijing Normal University

This paper presents a novel two-stage method for detecting structural changes in tensor sequences by leveraging Tucker decomposition to simulta- neously capture low-rank and sparse structures. In the first stage, we reformu- late the problem of detecting changes in the basis matrix sequence of the pro- jections as one of identifying discrete rank changes in a sequence of subspace ranks. This is achieved via an eigen-decomposition-based estimation proce- dure. In the second stage, we detect changes in the core tensor sequence and introduce a "Completion Algorithm" to handle core tensors of varying sizes across segments, enhancing the robustness of detection. This paper achieves computational efficiency through precise change point categorization and reduced redundancy. Theoretical guarantees include fast convergence rates of the estimated basis matrices and the consistency of the estimated locations and number of changes. Extensive experiments demonstrate the efficacy of the proposed approach in analyzing dynamic high-order tensor data. We further illustrate the method with a detailed analysis of a real-world dataset.

Low-Rank Spatio-Temporal State Space Models with Structured Spatial Covariance

◆ *Mengying You, Yuedong Wang, Wensheng Guo*

Shanghai University of International Business and Economics,University of California - Santa Barbara, University of Pennsylvania

We propose a low-rank spatio-temporal state space model for analyzing COVID-19 dynamics at the U.S. county level. The model incorporates a separable covariance structure to capture spatial and temporal dependence, where both components are estimated nonparametrically. This yields a smooth and stable low-rank representation of spatial variation.

The resulting state space formulation enables efficient estimation and accommodates shared temporal patterns as well as county-specific deviations. Applied to U.S. county-level data, the proposed method effectively captures spatial diffusion and improves reconstruction accuracy compared to approaches based on empirical covariance. The results demonstrate the importance of incorporating structured spatial dependence in modeling large-scale epidemic data.

S102: Recent Advances in Modeling Survival Data

Consistent Covariance Estimation for Stratum Imbalances Under Minimization Method for Covariate-Adaptive Randomization

Zixuan Zhao, Yanglei Song, ◆ *Wenyu Jiang, Dongsheng Tu*

Queen's University, Queen's University, Queen's University, Queen's University

Pocock and Simon's minimization method is a popular approach for covariate-adaptive randomization in clinical trials. Valid statistical inference with data collected under the minimization

method requires the knowledge of the limiting covariance matrix of within-stratum imbalances, whose existence is only recently established. In this work, we propose a bootstrap-based estimator for this limit and establish its consistency, in particular, by Le Cam's third lemma. As an application, we consider in simulation studies adjustments to existing robust tests for treatment effects with survival data by the proposed estimator. It shows that the adjusted tests achieve a size close to the nominal level, and unlike other designs, the robust tests without adjustment may have an asymptotic size inflation issue under the minimization method.

Efficient Marginal Estimation for Clustered Survival Data with Informative Correlation Structure

Cheng Yang, Changcheng Li, Song Leng, Yingwei Peng, ◆ *Yi Niu*

Dalian University of Technology, Dalian University of Technology, Dalian Medical University, Queen's University, Dalian University of Technology

Modeling clustered failure time data has been becoming increasingly important in biomedical studies. It is crucial to characterize the correlation among clustered failure times to enhance the efficiency of the estimation. Existing methods often employ random effects with known distributions or pre-specified working correlation structures to depict the dependence for clustered observations. In this talk, we propose a nonparametric method to estimate the correlation structure of clustered right-censored survival times based on the marginal Cox's proportional hazards model. We obtain an informative correlation structure by using the eigenvector basis matrices approximation. The number of basis matrices is determined by minimizing a discrepancy function with the SCAD penalty. The proposed method is able to examine the correlation patterns within clusters and avoid model misspecification. We show that the proposed method has the oracle properties and consistently selects the true correlation structure. We evaluate the performance of the proposed method through a simulation study and provide an application to the tooth loss data analysis.

A Proportional Hazards Mixture Cure Model for Subgroup Analysis: Inferential Method and an Application to Colon Cancer Data

◆ *Kai Liu, Yingwei Peng, Narayanaswamy Balakrishnan*

Shanghai Lixin University of Accounting and Finance, Queen's University, McMaster University

When determining subgroups with heterogeneous treatment effects in cancer clinical trials, the threshold of a variable that defines subgroups is often pre-determined by physicians based on their experience, and the optimality of the threshold is not well studied, par- ticularly when the mixture cure rate model is considered. We propose a mixture cure model that allows optimal subgroups to be estimated for both the time to event for uncured subjects and the cure status. We develop a smoothed maximum likelihood method for the estimation of model parameters. An extensive simulation study shows that the proposed smoothed maximum likelihood method provides accurate estimates. Finally, the proposed mixture cure model is applied to a colon cancer study to evaluate the potential differences in the treatment effect of levamisole plus fluorouracil therapy versus levamisole alone therapy between younger and older patients. The model suggests that the difference in the treatment effect on the time to cancer recurrence for uncured patients is significant between patients younger than 67 and patients older than 67, and the younger patient group benefits more from the combined therapy than the older patient group.

Average Treatment Effect Estimation for Survival Data with a Cure Fraction

Hang Liu, ◆ *Yingwei Peng*

University of Electronic Science and Technology of China, Queen's University

Estimating the causal effect of a continuous treatment on survival data, particularly in cases where there is a cured fraction from observational studies, is a significant issue. However, this topic is not well addressed in the existing literature. Current methods either rely on strong parametric assumptions or struggle to effectively control for confounding variables. In this study, we propose a novel nonparametric estimation method that utilizes a weighted generalized Kaplan-Meier survival estimator. This method aims to estimate the average effects of a continuous treatment on both the probability of being cured and the restricted mean survival time. Notably, our approach does not require any parametric assumptions about the effects, and it can efficiently control for multiple confounding variables. A simulation study demonstrates that our proposed method outperforms existing approaches, particularly when the average effects are complex or when confounding is strong. We apply this method to data from a study of chlamydia patients to evaluate the average effects of years of schooling on the probability of being immune to reinfection, as well as on the restricted mean survival time.

S108: Recent Advances in Spatio-Temporal Data Analysis

On Lasso and Other Variable Selection Algorithms with Moderately Dense Coefficients

Ching-Kang ING, Yan-Shuo PAN, ◆ *CY (Chor-yiu) SIN*

Tsing Hua University, Taiwan, Tsing Hua University, Taiwan, Tsing Hua University, Taiwan

In their seminal paper, Negahban, Ravikumar, Wainwright and Yu (2012, Statistical Science) showed the error bound of the Lasso estimator attains the minimax optimal rate when the number of candidate variables p in a high-dimensional linear model satisfies log p = o(n) and the corresponding regression coefficient lie in an L(q) ball of radius R(q), $0 < q < 1$, where n is the sample size. In this paper, we first devise a variable selection algorithm which also consider cases of so-called "moderately dense"coefficients, in which $1 < q < 2$.Second, we allow for non-decaying coefficients, and thus our algorithm can be applied to reduced-form equations in which the instruments are possibly weak. See, for instance, Mikusheva and Sun (2022, Rev. Econ. Stud.), and Mikusheva (2026, Book Chapter). Third, for cases of hard sparsity, depending on q, we allow for a wider range of number of non-zero coefficients s. Our algorithm is compared with Lasso via Monte-Carlo simulations, the latter of which is required, for $1 < q < 2$, that p is much more restrictive, by Sin, Pan and Ing (2026, TR).

MDD-Based Estimation for Multivariate Spatial Temporal Models with Heterogeneous Coefficients

Muyi Li

Xiamen University

This paper proposes an MDD-based estimator for multivariate spatio-temporal models with heterogeneous coefficients. The proposed framework extends the applicability of MDD-based estimation by accommodating spatial dependence, coefficient heterogeneity, and settings in which both the cross-sectional and time dimensions diverge. Compared with traditional moment-based methods for spatio-temporal models, the MDD-based estimator exploits information implied by conditional independence rather than relying solely on uncorrelatedness, and is therefore expected to achieve higher efficiency. The paper establishes the corresponding asymptotic theory. Simulation studies and a real data analysis demonstrate the finite-sample performance and application of the method.

A Double Self-Normalized Approach to Change-Point Detection

Yaxing Yang

Xiamen University

This paper proposes a novel test for the change point in time series based on a double self-normalized approach, which is distribution-free and involves no tuning parameters. The asymptotic properties of the proposed test are established under the null and alternative hypotheses. Testing for parameter constancy in a general time series regression setting is also considered. Monte Carlo simulations demonstrate the merits of our approach and illustrations using real data examples are presented.

Model Selection under Misspecification and Endogeneity: The MRIC Approach to Social Interaction Models

Hon Ho Kwok

Taiwan University

In this paper, we argue that model selection (MS) is particularly important in applying social interaction models by proving two theorems. First, suppose the true model is a system of n regressions of heterogeneous mean functions, then we can easily construct social interaction models that are more predictive than a system of n regressions of homogeneous mean functions. The spurious social effects can be interpreted as James-Stein shrinkage effects. Second, for a given social interaction model, there can be many social interaction models that are equivalent or close to it, in terms of prediction error. Hence, we establishes predictive-efficient MS criteria (and consistent nonlinear least squares (NLS) estimators) for a finite set of parametric, potentially misspecified, and non-nested social interaction models. The models under consideration include various forms of linear and nonlinear regression, social interaction, and spatial models. We develop new methods for consistently estimating penalties with general covariance or network structure in conditional means and regression errors (which are applicable even when all models are misspecified). The MS criteria are asymptotically efficient whether the true model is included. If there are candidate models which correctly specify the conditional mean, the most predictive one among them will be chosen with probability converges to 1. A discussion on MS consistency is provided.

S110: Recent Advances in Statistical Learning and Big Data

Variable Selection for Nadaraya-Watson-Frechet Regression

Yushen Dong, Youhan Lu, ◆ Yichao Wu, Hans-Georg Muller

University of Illinois Chicago, University of Illinois Chicago, University of Illinois Chicago, University of California at Davis

In this paper, we propose a new variable selection method for Nadaraya-Watson- Frechet regression, which is a special case of local Frechet regression. We demonstrate both the practical performance and the theoretical properties of this novel variable selection method. To provide theoretical support, we first derive a refined uniform convergence result for the Nadaraya-Watson-Frechet regression estimator of the Frechet regression function, which then enables us to establish asymptotic selection consistency for the proposed variable selection method. Excellent performance of the proposed variable selection method for Nadaraya-Watson-Frechet regression in finite sample situations is demonstrated both with simulation studies and in real data

examples.

Cluster Analysis of Longitudinal Associations Between Microbial Abundance and Environmental Factors

Emily Palmer, Caihong Qin, ◆ Lan Xue, Yuan Jiang

HP Inc., Indiana University, Oregon State University, Oregon State University

Microbial communities are dynamic systems that shift in composition in response to host and environmental factors. Understanding these temporal patterns is essential for identifying taxa with shared biological roles. However, standard clustering approaches for microbiome data often neglect its longitudinal structure and fail to account for external variables that may influence microbial trajectories. We introduce a novel clustering framework named \textit{cleverly}, denoting (Cl)ustering with (E)xternal (V)ariabl(e)s in (R) with (L)ongitudinal (Y)s, which incorporates external variables into the clustering of multivariate responses represented as longitudinal and compositional counts. Our method enables the generation of clusters based either on the responses' baseline trajectories or on their reactions to external panerturbations, as well as accommodates the compositional, discrete, and time-correlated nature of the data. Using simulated data, we demonstrate that \textit{cleverly} recovers true groupings under a range of biologically realistic conditions, including varying correlation structures and compositional constraints. We further apply \textit{cleverly} to longitudinal fecal microbiome data from wild African buffalo (\textit{Syncerus caffer}), revealing distinct clusters of taxa that either maintain consistent trajectories or respond sharply to changes in resource availability. These findings underscore the importance of accounting for environmental covariates in microbiome clustering and illustrate how \textit{cleverly} enables a deeper understanding of microbial function and host-microbiome interactions in dynamic ecosystems.

Multiply-Penalized Decentralized Estimation with an Application to Convex Support Tensor Machines

Bei Zhang, Zemin Zheng, ◆ Heng Lian

University of Science and Technology of China, City University of Hong Kong, University of Science and Technology of China, City University of Hong Kong Shenzhen Research Institute

We consider the penalized estimation problem involving multiple convex penalties in a decentralized setting where data are distributed on multiple nodes that are connected via a network. We propose an ADMM-based algorithm for efficient distributed computing and establish its convergence. The technical difference, compared to related works on decentralized ADMM, is in dealing with multiple penalties (distributed convergence properties using multiple proximal steps is an open problem as far as we know). An important example in focus is the convex support tensor machine (STM) for classification based on multiple nuclear norm penalties, which brings further challenges in analysis due to the non-smooth empirical loss, for which strong convexity holds only approximately (up to an approximation/statistical error term related to approximating the population loss using the empirical loss) and only within a bounded compact set of the Euclidean space. We illustrate the performance of the algorithm for STM using some numerical examples.

S113: Recent Advances on Statistical Inference and Econometrics

Factor-Adjusted Multiple Testing for High-Dimensional Individual Mediation Effects

Chen Shi, ◆ Zhao Chen, Dan Wang

Fudan University, Fudan University, New York University Shanghai

Understanding the mechanisms through which a treatment or exposure affects an outcome is a central goal in mediation analysis. In high-dimensional settings with numerous mediators, identifying individual mediation effects is challenging due to complex dependence structures and the need to control the false discovery rate (FDR). This paper develops a novel factor-adjusted inference framework for testing high-dimensional individual mediation effects, explicitly accounting for strong correlations among mediators. By modeling the error terms of the mediator model via an approximate factor model, we extract latent factors that capture dependence and use the resulting decorrelated residuals as pseudo-predictors for valid inference. We establish the asymptotic normality of the debiased estimator and develop a procedure to control the FDR in multiple testing under mild assumptions. Our approach not only improves statistical reliability but also enhances robustness to latent confounding, which is critical in observational studies. Extensive Monte Carlo simulations demonstrate the superior finite-sample performance of the proposed method in terms of both statistical power and FDR control. We further illustrate its practical utility through a real-data application to multi-omics breast cancer data from TCGA.

Estimation of Rough Leverage Effect

Ming Cheng, ◆ Christina Dan Wang,Ziyang Xiong

Fudan University, New York University Shanghai, Fudan University

This paper studies high-frequency estimation of leverage effects under rough volatility. The different local scales of return and

rough variance increments make the classical semimartingale leverage normalization inappropriate. We introduce a scale-adapted rough leverage functional, derive oracle and feasible asymptotic theories, and show that the estimator exhibits a phase transition driven by the Hurst parameter and the structure of the rough volatility coefficient. For feasible inference, local realized-variance estimation introduces an additional smoothing effect that changes the rate and asymptotic variance. Simulations show that the proposed observable variance estimator provides good finite-sample calibration across several rough volatility specifications. An empirical application to high-frequency U.S. equity and ETF data finds persistent negative rough leverage and substantial differences from the benchmark semimartingale normalization.

A General Model Averaging Framework with Inference: Nested or Non-Nested Models

Yufei Li, ◆ *Fang Fang, Hansheng Wang*

East China Normal University

As an important alternative to model selection for handling model uncertainty, frequentist model averaging is attractive for its competitive prediction power. However, it has long been criticized for the deffciency of statistical inference. The post-averaging inference methods of model averaging are quite limited especially for non-nested models. To ffll the gap, we propose a general model averaging framework with inference, which is applicable to linear model, generalized linear model, Cox model, generalized estimating equation and many others, and both nested and non-nested candidate models are considered. The asymptotic properties of the optimal weight, estimation consistency and asymptotic distribution of the model averaging estimator are established. A simulation-based inference procedure is proposed to construct the conffdence interval. Extensive simulation studies and a credit card data analysis conffrm the theoretical results and show the advantage of the proposed method for a wide range of problems.

Distributed Semi-Supervised Inference for Generalized Linear Models with Block-Wise Missing Covariates

Ziyuan Wang, Jin Liu, Jun Shao, Heng Lian, ◆ *Lei Wang*

Nankai University, Nankai University, University of Wisconsin-Madison, City University of Hong Kong, Nankai University

For a relatively small labeled dataset from high-dimensional generalized linear models with block-wise missing covariates and a large unlabeled dataset, we utilize a model-assisted approach in the labeled dataset to address the issue of block-wise missing covariates and then integrate the unlabeled data to construct estimation equations for the coefficients without any imputation. A lasso-penalized semi-supervised estimator is obtained, and then its debiased estimator is proposed to establish asymptotic normality/confidence intervals. When the labeled data are distributed in multiple machines independently and only some machines have unlabeled data, we further propose a distributed debiased semi-supervised estimator for estimation and inference. The finite sample performance of our proposed two estimators is studied through simulations and further illustrated with a breast cancer dataset.

S116: Recent Advances in Complex Time Series and Dynamic Network Analysis

Nonlinear Independent Component Analysis for Time Series

Han Yan

London School of Economics and Political Science

Independent component analysis (ICA) is a fundamental tool for dimension reduction and recovering latent independent signals from multivariate observations with broad applications across various fields. Most existing ICA methods, however, are designed for linear mixtures and independently observed data, and are therefore inadequate for nonlinearly mixed data or the data exhibiting temporal dependence. To fill in the gap, we propose a nonlinear ICA for multivariate time series such that the transformed time series are independent. contemporaneously and serially. We use an energy-distance to measure independence and employ invertible neural networks (INN) with coupling flows to represent nonlinear and invertible transformations. We establish theoretical guarantees by characterizing the approximation capacity and function class complexity of the INN class and deriving a convergence rate for the nonlinear ICA estimator. Numerical experiments demonstrate that the proposed method performs well in decomposing multivariate time series into mutually independent component series.

Change Point Detection and Identification in Tensor Factor Models

Zetai Cen
University of Bristol

We study multiple change point detection in a high-dimensional tensor factor model based on Tucker decomposition, where the post-segmentation change identification is also thoroughly investigated, which is new to the literature. We formalize the identification concept such that each detected change could be attributed to some tensor modes and hence called mode-identifiable, whereas a change is mode-unidentifiable if no modes are assigned. Our detection algorithm employs the narrowest-over-threshold principle to search change points on seeded intervals, and hence can effectively localize non-linearly spaced

change points at near-linear computational cost. Theoretical results on the localization consistency are provided, and we give necessary and sufficient conditions on the types of changes that are mode-identifiable. We also propose a method to identify the modes corresponding to each localized change, with consistency guaranteed. Numerical results show that our procedures in both detection and identification work well.

Hypergraphs Embeddings: A Novel Approach with Increasing Dimensions

Xinyang Yu
London School of Economics and Political Science

We propose the inherited embedding method within the hypergraph framework. Unlike traditional representation methods that use a fixed dimension, our approach embeds hypergraphs into an extended Euclidean space, where the dimension increases with the order of hyperedges. We begin by introducing the framework of the inherited embedding model. A thorough study of the model and parameter estimation within the penalized maximum likelihood framework has been conducted. We also design a minorization-maximization (MM) algorithm for the numerical computation of parameter estimation. Extensive simulations on synthetic data are performed to verify the robustness of our model and algorithm. Finally, we apply the method to various real-world network datasets, revealing more intriguing results compared to traditional embedding methods.

Autoregressive Hypergraphs

Xianghe Zhu
London School of Economics and Political Science

We propose a first-order autoregressive (AR(1)) model for dynamic hypergraph processes that generalizes traditional network models to capture higher-order interactions among multiple nodes. In this framework, hyperedges evolve over time while the node set remains fixed. Building on this, we introduce an AR(1) stochastic block model for hypergraphs in which latent communities are characterized by time-varying transition probabilities. To uncover these communities, we develop a new spectral clustering algorithm based on the hypergraph Laplacian. We also integrate change-point inference into the AR(1) hypergraph stochastic block model to detect potential structural shifts. Finally, applications to two real-world datasets demonstrate the practical value and effectiveness of both the model and its associated inference techniques.

S120: Robust and Flexible Statistical Methods for Complex and High-Dimensional Data

Robust Multi-Task Learning for Principal Component Analysis

Dali Liu, ◆ *Haolei Weng*

Michigan State University, Southern University of Science and Technology

Principal component analysis (PCA) is a fundamental tool in statistical learning. For data collected from multiple sources, we propose new multi-task learning PCA algorithms that can adaptively leverage the unknown distribution similarities to enhance eigenspace estimation and are robust to data contamination. The first proposed algorithm is shown to achieve (near) minimax optimal rate for the average estimation error. Leveraging a notion of matrix depth, we then develop a second algorithm that is minimax optimal for the maximum estimation error under certain conditions. Numerical studies are provided to demonstrate the effectiveness of the proposed methods.

Double Penalized Variable Selection for High-Dimensional Partial Linear Mixed Effects Models

Yiping Yang, Chuanqin Luo, ◆ *Weiming Yang*

Nanjing University Of Finance & Economics, Chongqing Technology and Business University, Chongqing Technology and Business University

In this study, we address the selection of both fixed and random effects in partial linear mixed effects models. By combining B-spline and QR decomposition techniques, we propose a double-penalized likelihood procedure for both estimating and selecting these effects. Furthermore, we introduce an orthogonality-based method to estimate the non-parametric component, ensuring that the fixed and random effects are separated without any mutual interference. The asymptotic properties of the resulting estimators are investigated under mild conditions. Simulation studies are conducted to evaluate the finite sample performance of the proposed method. Finally, we demonstrate the practical applicability of our methodology by analyzing a real data.

Functional Additive Model via Convolution Smooth Quantile Estimation

◆ *Ping Yu, Xiaojing Liu*

Shanxi Normal University, Shanxi Normal University

Advancements in modern science have led to an increased prevalence of functional data, which are usually viewed as elements of the space of square-integrable functions L^2. This paper considers the convolution-smoothed quantile estimation and variable selection problem for functional additive model. We study the convolution-smoothed quantile estimation of a functional additive model constructed for a scalar response variable based on scores of functional predictors. Firstly, the nonparametric effects of the functional principal component scores are modeled as additive components, and the relevant

principal components are selected using the group LASSO method, with B-splines employed for approximation. Second, convolution--smoothed quantile regression is applied to obtain estimators of the spline coefficient vector and the nonparametric functions. Under certain regularity conditions, we establish the convergence rates and smoothing bias of the spline coefficient vector and the nonparametric function estimators. Finally, simulation studies and applications to the Tetouan dataset and Beijing multi-site air quality data demonstrate the effectiveness of the proposed method.

Model-Free Rank Aggregation: A Maximum Score Approach

◆ *Haoran Zhang, Yunxiao Chen*

Southern University of Science and Technology, London School of Economics and Political Science

This paper addresses the rank aggregation problem using multi-way comparison data derived from raters' scores. Unlike traditional methods that rely on parametric models (e.g., Bradley-Terry and Plackett-Luce models), we consider a model-free setting that assumes only the existence of a latent global ranking and a weak monotonicity constraint for each rater. This flexible setting accounts for heterogeneity in raters and also covers the weak stochastic transitivity setting for pairwise comparison data as a special case. To learn this global ranking, we introduce a distribution-free maximum score estimator that automatically accounts for rater heterogeneity. This estimator is effective across both dense and sparse regimes, accommodating scenarios where the number of ratings per rater is either a fixed constant or diverges with the total number of items. We establish the theoretical foundations of the proposed estimator by proving its consistency, showing that the proportion of discordant pairs (Kendall's tau) converges to zero in probability as the number of raters diverges. Furthermore, we derive tight upper and lower bounds for a performance metric based on Kendall's tau. We show that under specific asymptotic regimes, these bounds match, establishing the minimax optimality of our method. The practical utility of the estimator is demonstrated through extensive simulations and two real-data applications, one on ranking sport players and the other on aggregating preferences of survey respondents.

S125: Statistical and Computational Advances for Single-Cell Genomics

High-Resolution Reconstruction of Single-Cell Specific Spatial Genome Architectures in 3D Space

Jianrong Wang

Michigan State University

Spatial chromosomal conformations in 3D space play pivotal roles in modulating interactions of long-range transcription regulation and demonstrates high levels of cell-to-cell heterogeneity. However, current single-cell Hi-C experiments are highly limited to just a few cellular-contexts with extremely high rates of missing contacts (>99.9%), making high-resolution characterization of detailed genome folding challenging. We have developed a family of computational models (Tensor-FLAMINGO and its variants), which are able to reconstruct 10kb-resolution spatial configurations of the human and mouse genomes (i.e. the highest resolution to date) at single-cell specific levels. Tensor-FLAMINGO consistently demonstrate superior reconstruction accuracy and strong robustness against high rates of missing contacts. Integrative analysis of reconstructed single-cell 3D genome structures with context-specific multi-omics data and genetic association studies revealed a series of novel discoveries, including cell-specific spatial hubs of multi-way interactions, geometric properties of chromatin loops, interplay between epigenetic landscapes and 3D folding, long-range eQTLs (>900kb), and significant single-cell specific long-range regulatory interactions. The completion of missing chromatin contacts further enables the characterization of regulatory networks for the majority of genes, which is not possible based on the highly sparse experimental data. Beyond 1D genome annotations and 2D contact map analyses, our predictive and analytical framework of single-cell 3D conformation promotes systems-level insights into the spatial coordination of regulatory programs and its functional impacts, across diverse cellular-contexts and at unprecedented resolution.

CellScope: High-Performance Cell Atlas Workflow with Tree-Structured Representation

◆ *Bingjie Li, Runyu Lin, Tianhao Ni, Guanao Yan, Mannix Burns, Jingyi Li, Zhigang Yao*

The Shanghai Institute for Mathematics and Interdisciplinary Sciences, National University of Singapore, Zhejiang University, Michigan State University, University of California, Los Angeles, Fred Hutchinson Cancer Center, National University of Singapore

Single-cell sequencing enables comprehensive profiling of individual cells, revealing cellular heterogeneity and function with unprecedented resolution. However, current analysis frameworks lack the ability to simultaneously explore and visualize cellular hierarchies at multiple biological levels. To address these limitations, we present CellScope, a promising framework for constructing high-resolution cell atlases at multiple clustering levels. CellScope employs a two-stage manifold fitting process for gene selection and noise reduction, followed by agglomerative clustering, and integrates UMAP visualization with hierarchical clustering to intuitively represent cellular relationships simultaneously at multiple levels—such

as cell lineage, cell type, and cell subtype levels. Compared to established pipelines such as Seurat and Scanpy, CellScope comprehensively improves clustering performance, visualization clarity, computational efficiency, and algorithm interpretability, while reducing dependence on hyperparameters across a multitude of single-cell datasets. Most importantly, it can reveal biological insights that other contemporary methods are unable to detect, thereby deepening our understanding of cellular heterogeneity and function, and potentially informing disease research.

SpaDOT: Optimal Transport Modeling Uncovers Spatial Domain Dynamics in Spatiotemporal Transcriptomics Studies

◆ *Wenjing Ma, Siyu Hou, Lulu Shang, Jiaying Lu, Xiang Zhou*

The Chinese University of Hong Kong, Shenzhen, Yale University, The University of Texas MD Anderson Cancer Center, Emory University, Yale University

Spatiotemporal transcriptomics is an emerging and powerful approach that adds a temporal dimension to spatial transcriptomics, enabling the characterization of dynamic changes in tissue architecture during development or disease progression. Here, we present SpaDOT (Spatial Domain Transition detection), a computational method designed to identify spatial domains and infer their temporal dynamics across time points for spatiotemporal transcriptomics studies. SpaDOT employs a variational autoencoder (VAE) framework to capture a low-dimensional representation of the data, incorporating hidden clustering variables to define distinct spatial domains. In the process, SpaDOT integrates a Gaussian Process prior and graph neighbor information to explicitly model both global spatial continuity and local structural heterogeneity among spatial locations within each time point, while using optimal transport (OT) to derive time-varying embeddings and infer the relationships between spatial domains across time points. Through simulation and real data applications, we demonstrate the superiority of SpaDOT in spatial domain detection, latent space preservation, and domain transition tracking over time. In real data applications, SpaDOT accurately captures dynamic domain transitions, including the disappearance and re-emergence of domains due to technical variation.

Controlling False Discoveries After Clustering via Data Splitting for Spatial Domain Marker Detection

◆ *Yingxin Lin, Lijun Wang, Hongyu Zhao*

The Chinese University of Hong Kong, Zhejiang University, Yale University

Spatial omics technologies are transforming biomedical research by enabling genome-wide measurement of molecular activity while preserving the spatial context within tissues. These advances create unprecedented opportunities to uncover cell–cell interactions, tissue organization, and disease mechanisms. A crucial step in realizing this potential is identifying spatial domain markers, which are essential for defining tissue architecture and understanding disease progression. However, using the same data for both clustering and marker detection creates the problem of "double dipping", which can lead to inflated false discoveries, particularly when domain boundaries are poorly defined. To address this challenge, we develop SpaDS and SpaMDS, data splitting based approaches for differential expression testing after clustering for spatial omics, enabling robust spatial domain marker discovery with controlled false discovery rate and high power. Through extensive simulations and analyses of spatial omics datasets from diverse technologies, we demonstrate that the data-splitting methods are easy to implement, adaptable to existing spatial omics data analysis pipelines, and often outperform other approaches under weak signals and high feature correlations.

S134: Statistical Inference for Complex Data Structures

Estimating Causal Risk Ratio Under Complex Longitudinal Adherence Patterns with Time-Varying Principal Stratification

Rui Wang, Kwun Chuen Gary Chan, ◆ *Fei Gao*

University of Washington, University of Washington, Fred Hutchinson Cancer Center

Estimating causal treatment effects in the presence of time-varying adherence requires addressing post-randomization selection bias and unmeasured confounding. We introduce a time-varying principal stratification framework to tackle these challenges and define a causal risk ratio for time-to-event outcomes. To estimate this effect, we develop two approaches: a moment estimator and a maximum likelihood estimator. Our framework accommodates settings with partially observed adherence data by leveraging efficient sampling designs, including case-cohort and nested case-control schemes. We demonstrate the utility of our methods through applications to two HIV prevention studies.

Mediation with Random Effect Exposures

◆ *Fan Xia, Chuck McCulloch*

University of California, San Francisco, University of California, San Francisco

Causal mediation analysis is typically formulated using mean contrasts between a treatment and a reference condition, an approach that is ill-suited when the exposure has many levels and no natural control group. Such settings arise, for example, when the exposure represents a population of units such as providers or omics features and is naturally modeled as a random

effect. In these contexts, variance-based summaries such as percent variance explained are widely used but lack a causal interpretation. We develop a causal framework for variance-based mediation analysis with random-effect exposures, defining total, direct, and indirect effects through sequential interventions that respect causal ordering without relying on an arbitrary reference exposure. We establish identification under linearity and propose efficient estimation using joint mixed-effects models.

Simultaneous Clustering and Estimation of Additive Shape Invariant Models for Recurrent Event Data

Zitong Zhang, ◆ *Shizhe Chen*

University of California, Davis, University of California, Davis

Technological advancements have enabled the recording of spiking activities from large neuron ensembles, presenting an exciting yet challenging opportunity for statistical analysis. This project considers the challenges from a common type of neuroscience experiments, where randomized interventions are applied over the course of each trial. The objective is to identify groups of neurons with unique stimulation responses and estimate these responses. The observed data, however, comprise superpositions of neural responses to all stimuli, which is further complicated by varying response latencies across neurons. We introduce a novel additive shape invariant model that is capable of simultaneously accommodating multiple clusters, additive components, and unknown time-shifts. We establish conditions for the identifiability of model parameters, offering guidance for the design of future experiments. We examine the properties of the proposed algorithm through simulation studies, and apply the proposed method on neural data collected in mice. Specifically, we analyze neural spike trains recorded using Neuropixels probes during a visual discrimination task with two independent randomized stimuli. The analysis identifies three distinct functional groups of neurons that exhibit heterogeneous response patterns to the stimuli, including both stimulus-specific activity and variability in response timing.

Reluctant Interaction Inference after Additive Modeling

Yiling Huang, Snigdha Panigrahi, ◆ *Guo Yu, Jacob Bien*

University of Michigan, University of Michigan, University of California Santa Barbara, University of Southern California

Additive models enjoy the flexibility of nonlinear models while still being readily understandable to humans. By contrast, other nonlinear models, which involve interactions between features, are not only harder to fit but also substantially more complicated to explain. Guided by the principle of parsimony, a data analyst therefore may naturally be reluctant to move beyond an additive model unless it is truly warranted. To put this principle of interaction reluctance into practice, we formulate the problem as a hypothesis test with a fitted sparse additive model (SPAM) serving as the null. Because our hypotheses on interaction effects are formed after fitting a SPAM to the data, we adopt a selective inference approach to construct p-values that properly account for this data adaptivity. Our approach makes use of external randomization to obtain the distribution of test statistics conditional on the SPAM fit, allowing us to derive valid p-values, corrected for the over-optimism introduced by the data-adaptive process prior to the test. Through experiments on simulated and real data, we illustrate that—even with small amounts of external randomization—this rigorous modeling approach enjoys considerable advantages over naive methods and data splitting.

S152: Toward Efficient Clinical Trials: Statistical Methods That Leverage Baseline Covariates

To Adaptively Randomize or to Rerandomize: A Comparison of Covariate-Adaptive Randomization and Rerandomization

Ziji Qin, Feifang Hu, ◆ *Yang Liu*

The George Washington University, The George Washington University, Renmin University of China

In comparative studies, achieving balance in influential baseline covariates is fundamental to the credible assessment of treatment effects. Two prominent strategies for improving covariate balance are covariate-adaptive randomization (CAR) and rerandomization (RR). CAR was originally designed for settings in which subjects are enrolled sequentially, whereas RR requires that all subjects enter the trial simultaneously. Although both approaches aim to mitigate covariate imbalance, their underlying mechanisms differ substantially, which have hindered systematic comparisons and left important theoretical questions unresolved. This article develops a unified framework that enables a rigorous comparison of CAR and RR when both procedures target an identical imbalance criterion. Within this framework, we establish conditions under which RR achieves balance properties comparable to those of CAR and characterize how each design affects the asymptotic behavior of treatment effect estimators. We further investigate the computational burden of RR, showing that it grows linearly with the sample size but exponentially with the covariate dimension. As a result, while RR can be effective in small to moderate samples, its computational cost escalates rapidly, limiting its scalability in large trials or experiments with a large number of covariates. In contrast, CAR retains computational efficiency while delivering strong balance guarantees in both sequential and simultaneous enrollment settings.

Shape-Preserving Covariate Adjustment via Empirical Likeli-

hood in Randomized Experiments

◆ *Zhilan Lou, Jun Shao, Tuo Wang, Yanyao Yi, Yu Du, Ting Ye*

Zhejiang University of Finance and Economics, University of Wisconsin - Madison, Eli Lilly and Company, Eli Lilly and Company, Eli Lilly and Company, University of Washington

Covariate adjustment improves estimation efficiency in randomized experiments, but standard calibration and augmentation methods, when applied to distribution or survival functions, do not preserve monotonicity—a fundamental property of the estimand. We propose using empirical likelihood with covariate-balancing constraints to construct a covariate-adjusted empirical measure for each treatment arm. Estimators of a broad class of distributional functionals—including cumulative distributions, survival functions, quantiles, and area-under-the-curve treatment effects—are then derived as plug-in functionals of this measure, automatically inheriting proper shape constraints. We establish asymptotic normality with an explicit, guaranteed efficiency gain over unadjusted estimators. The asymptotic distributions are invariant to the randomization scheme, providing a unified inference procedure under simple randomization and all commonly used covariate-adaptive designs satisfying a mild balancing condition. This unified construction—adjust the empirical measure once, then derive any estimator from it—provides a principled reconciliation of covariate adjustment with shape preservation in randomized experiments. Simulations and an application to the SURPASS-4 trial confirm the theoretical gains.

A Robust Score Test in G-Computation for Covariate Adjustment in Randomized Clinical Trials

◆ *Xin Zhang, Haitao Chu, Lin Liu, Satrajit Roychoudhury*

Pfizer Inc., Pfizer Inc., Shanghai Jiao Tong University, Pfizer Inc.

G-computation has become a widely used robust method for estimating unconditional (marginal) treatment effects with covariate adjustment in the analysis of randomized clinical trials. Statistical inference in this context typically relies on the Wald test or Wald interval, which can be easily implemented using a consistent variance estimator. However, existing literature suggests that when sample sizes are small or when parameters of interest are near boundary values, Wald-based methods may be less reliable due to type I error rate inflation and insufficient interval coverage. In this article, we propose a robust score test for g-computation estimators in the context of two-sample treatment comparisons. The proposed test is asymptotically valid under simple and stratified (biased-coin) randomization schemes, even when regression models are misspecified. These test statistics can be conveniently computed using existing variance estimators, and the corresponding confidence intervals have closed-form expressions, making them convenient to implement. Through extensive simulations, we demonstrate the superior finite-sample performance of the proposed method. Finally, we apply the proposed method to reanalyze a completed randomized clinical trial. The new analysis using our proposed score test achieves statistical significance, whilst reducing the issue of type I error inflation.

Inverse Probability of Treatment Weighting: A Simple and Effective Approach to Covariate Adjustment for Survival Endpoints in Randomized Clinical Trials

Yongwu Shao, ◆ *Zhiwei Zhang, Zhishen Ye*

Gilead Sciences Inc., Gilead Sciences Inc., Gilead Sciences Inc.

Covariate adjustment aims to improve the statistical efficiency of randomized trials by incorporating information from baseline covariates. Popular methods for covariate adjustment include analysis of covariance for continuous endpoints and standardized logistic regression for binary endpoints. For survival endpoints, while some covariate adjustment methods have been developed, they are not commonly used in practice for various reasons, including high demands for theoretical and methodological sophistication as well as computational skills. In this talk, we point out that inverse probability of treatment weighting (IPTW) is a simple and effective approach to covariate adjustment for survival endpoints. We provide theoretical results that justify its use in conjunction with the Cox model and the Kaplan-Meier analysis, as well as numerical results that demonstrate its effectiveness in finite samples of small to moderate sizes. Compared to other covariate adjustment methods for survival endpoints, IPTW has several major advantages including simplicity, generality and ease of implementation, and thus holds great promise as a practical approach to covariate adjustment for survival endpoints.

S074: Modern Statistical Methods for Complex Longitudinal Data and Causal Inference

Longitudinal Mixed Membership Image-on-Scalar Model

Zhihao Wu, ◆ *Xinyuan Song, Kai Kang*

Yale University, The Chinese University of Hong Kong, Sun Yat-sen University

Magnetic resonance imaging (MRI) data have been extensively applied in diagnosing and predicting Alzheimer's disease (AD). However, there has been a notable oversight in addressing individual heterogeneity within longitudinal MRI data. This study introduces a novel modeling framework to elucidate the diverse dynamic patterns inherent in longitudinal imaging data, thereby facilitating a better understanding of individualized AD progression. The framework commences with a basis expansion approach to approximate the longitudinal images. Subsequently, a vector of probability weights is introduced,

delineating a subject's partial membership across clusters. Such partial membership allows the subject's repeatedly measured imaging data to belong to different clusters. Finally, a nonlinear trajectory model is employed to capture the typical normal aging process and the potential transition from normal to severe stages during the disease course. A Bayesian approach coupled with efficient MCMC algorithms is developed for statistical inference. Extensive simulation results demonstrate the efficacy of the proposed methods in parameter estimation and model selection. The framework is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, yielding insights into the evolving patterns of brain structures throughout AD progression.

A Hierarchical Gaussian Process Framework for Longitudinal Ordinal Digital Survey Responses

Zoe Gibbs McBride, ◆ *Xiaojing Wang*

Brigham Young University,University of Connecticut

Longitudinal mobile assessment studies often collect repeated multi-item survey responses over time. These data are useful for studying changes in self-reported well-being, but they can be challenging to analyze because responses are typically ordinal and collected at unequal time intervals. In this paper, we develop a flexible statistical modeling framework for longitudinal ordinal data. The proposed approach combines Gaussian processes with hierarchical modeling ideas inspired by item-response theory, allowing individual response patterns to be represented as smooth trajectories over time. A participant's repeated survey responses can be viewed as a high-dimensional longitudinal observation, and the Gaussian process framework provides a natural way to capture both individual-level variation and overall temporal trends. We also incorporate basis splines to model nonlinear mean patterns and use a Hilbert space approximation to improve computational efficiency for Bayesian inference. The method is illustrated using data from a longitudinal mobile survey study of self-reported well-being among young adults.

A Multiply Robust Estimation in Learning Optimal Individualized Treatment Regimes

Wensheng Zhu

Yunnan University

One of the fundamental aspects of precision medicine is the identification of an optimal individualized treatment regimen, which aims to recommend treatment decisions that maximize the expected outcome for each patient based on their unique characteristics. Traditionally, doubly robust methods provide double protection for estimation consistency against model misspecifications. However, they allow only a single model for the propensity score and a single model for the outcome, and the assumption that one of these two models is correctly specified is restrictive in practice. To further enhance the robustness of the estimates based on doubly robust estimators, we propose a value function estimator with multiply robust by constructing multiple models for both the propensity score and the outcome variable. We prove that the estimator remains consistent as long as at least one model in the candidate set of the propensity score model or the candidate set of the outcome model is correctly specified. We further demonstrate that if both the candidate sets of the propensity score model and the outcome model contain at least one correctly specified model, the proposed estimator is asymptotically normal and achieves the semiparametric efficiency bound. Extensive simulation studies underscore the superior performance of our method in comparison to existing approaches. The application of the proposed method is further illustrated through analysis of data from the China Rural Hypertension Control Project.

A Unified Mendelian Randomization Framework for Inferring Causal Relationships: Accounting for Measured Covariates and Unmeasured Confounders

Xianghong Hu

Shenzhen University

Mendelian randomization (MR) is a powerful approach for causal inference, however, its reliability is frequently compromised by unadjusted covariates and unmeasured confounders, such as unmeasured pleiotropy and sample structure. To address these challenges, we introduce MRMU,

a novel paradigm for the MR framework. Unlike traditional single-variable or multivariable MR methods, MRMU selects instrumental variables only from the exposure of interest and

estimates one exposure effect at a time, while jointly accounting for measured covariates and unmeasured confounders. This design improves the reliability of MR analyses. In simulations

and real data, MRMU achieved improved type I error control, higher statistical power, and more accurate effect estimation than existing MR methods. Applying to coronary artery disease (CAD), MRMU identified robust cardiometabolic risk factors, including LDL-C, APOB, systolic blood pressure, body mass index, and smoking initiation, with consistent evidence across multiple CAD datasets. In contrast, traits such as HDL-C, height, and educational attainment, which were significant in existing MR methods, were no longer supported by MRMU. MRMU further supported blood pressure-related traits, rather than lipid traits, as the more relevant pathway linking urate to CAD. Finally, by integrating largescale plasma proteomics data, MRMU identified candidate CAD drug targets beyond established HMGCR and PCSK9-related pathways, highlighting its utility for therapeutic target prioritization.

Southern University of Science and Technology

Southern University of Science and Technology (SUSTech) is an innovation-oriented public university founded by Shenzhen government in the background of China's higher education reform. It aspires to be a model and pioneer for promoting higher education reform. It is committed to serving the mission of promoting Shenzhen as a modern, international, and innovative city and China as a creative country. On February 14, 2022, the Ministry of Education and two other national authorities released the list of universities and disciplines included in the second round of the Double World-Class Project. SUSTech and its mathematics discipline were selected onto the list. Drawing on the disciplinary structures and operational models of world-class universities of science and engineering, SUSTech features a disciplinary system focused on science, engineering, and medicine, complemented by business education and distinctive humanities and social sciences. It offers academic programs at undergraduate, master's, and doctoral levels, and conducts cutting-edge research across a range of emerging disciplinary areas, positioning itself as a think tank for social progress and a source of new knowledge and technologies.

Rooted in China, SUSTech has seized the historic opportunities brought by the dual drivers of the Guangdong-Hong Kong-Macao Greater Bay Area and the Pilot Demonstration Zone of Socialism with Chinese Characteristics, as well as the overlapping advantages of the Shenzhen Special Economic Zone and the Pilot Demonstration Zone. Upholding its founding spirit of "Daring to Explore, Pursuing Truth and Practicality, Advancing Reform and Innovation, Striving for Excellence", highlighting its educational characteristics of "Knowledge Creation, Innovation, Entrepreneurship", and practicing its motto "Virtue, Truth, Daily Renewal and Self-Strengthening", SUSTech is committed to serving the development of an innovation-oriented country and Shenzhen's goal of becoming an international, modern, innovation-driven city. It is rapidly growing into a world-class, high-level international research university that gathers top-tier faculty, nurtures outstanding innovative talents, produces world-leading academic achievements, and promotes the application of scientific and technological advances, laying a solid foundation for realizing its grand vision of becoming a world-class research university at an early date.

To date, SUSTech has signed agreements with approximately 1,300 faculty members, including 57 academicians, 68 international fellows, 57 recipients of the National Science Fund for Distinguished Young Scholars, and 42 recipients of the National Science Fund for Excellent Young Scholars. More than 90% of its teaching and research faculty have overseas work experience, over 60% have worked or studied at universities ranked among the world's top 100, and high-level talents account for more than 50% of the teaching and research faculty.

The university is committed to establishing a diversified admission mechanism conducive to educational equity and the selection of innovative talents, as well as an advanced comprehensive education system. It has taken the lead in reforming its admission system by adopting an admission model based on comprehensive evaluation with gaokao (national college entrance examination) as a key component. Centered on the "credit system, college system, and tutorial system" as well as the principles of "internationalization, individualization, and elite education", SUSTech vigorously cultivates top innovative talents. The university adopts a residential college management model and currently has six residential colleges: Zhiren College, Shuren College, Zhicheng College, Shude College, Zhixin College, and Shuli College.

As a young university, in line with its positioning and goals as a high-level international research university, SUSTech responds to the major strategic needs of national emerging industries and those of the Guangdong-Hong Kong-Macao Greater Bay Area. With a disciplinary structure focused on science, engineering, and medicine, complemented by business and distinctive humanities and social sciences, it has swiftly established a complete talent training system covering

bachelor's, master's, and doctoral programs.

SUSTech has established ten academic schools, including the School of Science and the School of Engineering, with 33 departments and several research centers, offering 41 undergraduate majors. In 2018, SUSTech was accredited as a doctoral and master's degree conferring institution. Since its establishment, it has been granted 14 first-level doctoral programs, 3 professional doctoral programs, 15 first-level master's programs, 8 professional master's programs, and 7 postdoctoral research stations.

The university covers a total land area of 1.98 million square meters. The planned total gross floor area is 1.7162 million square meters, of which 975,000 square meters have been completed. A number of new buildings in Phase II of the campus, including the School of Engineering, School of Science, School of Business, and the Third Teaching Building, have been fully put into use. Construction of Phase III projects, including the SUSTech Affiliated Hospital (Main Campus), School of Medicine, and faculty housing, is progressing steadily and will be completed and delivered successively in 2026.

In addition, SUSTech has incubated Shenzhen Ocean University, whose construction has commenced with a total floor area of 600,500 square meters.

Boasting a beautiful environment and convenient transportation, the campus features nine hills and one water body, presenting picturesque scenery. The campus architecture embodies the principles of "substantiality, energy efficiency, practicality, and environmental protection", making it a model of green buildings and ecological campuses. SUSTech strives to become one of the most beautiful university campuses in China.

南方科技大学

南方科技大学（简称：南科大）是深圳在中国高等教育改革发展的时代背景下创建的一所高起点、高定位的公办新型研究型大学。2022 年 2 月 14 日，教育部等三部委公布第二轮“双一流”建设高校及建设学科名单，南方科技大学及数学学科入选“双一流”建设高校及建设学科名单。学校借鉴世界一流理工科大学的学科设置和办学模式，以理、工、医为主，兼具商科和特色人文社科的学科体系，在本科、硕士、博士层次办学，在一系列新的学科方向上开展研究，使学校成为引领社会发展的思想库和新知识、新技术的源泉。

南方科技大学扎根中国大地，紧抓粤港澳大湾区、深圳先行示范区“双区”驱动，深圳经济特区、深圳先行示范区“双区”叠加的历史机遇，发扬“敢闯敢试、求真务实、改革创新、追求卓越”的创校精神，突出“创知、创新、创业”的办学特色，践行“明德求是、日新自强”的校训精神，努力服务创新型国家建设及深圳国际化现代化创新型城市建设，快速建设成为聚集一流师资、培养拔尖创新人才、创造国际一流学术成果并推动科技应用的国际化高水平研究型大学，为尽早实现建成世界一流研究型大学的宏伟目标打下坚实基础。

截至目前，南方科技大学已签约引进教师约 1300 人，包括院士 57 人，国际会士 68 人，“国家自然科学基金杰出青年基金”获得者 57 人、“国家自然科学基金优秀青年基金”获得者 42 人。教学科研系列教师 90% 以上具有海外工作经验，60% 以上具有在世界排名前 100 名大学工作或学习的经历，高层次人才占教研系列 50% 以上。

学校着力建立有利于教育公平、创新人才选拔的多元录取机制和先进的全面教育体系。学校率先改革招生制度，采用“基于高考的综合评价录取模式”招收优秀学生。学校以“学分制、书院制、导师制”和“国际化、个性化、精英化”为核心和特色，大力培养拔尖创新人才。学校采用书院制管理，目前共设立致仁、树仁、致诚、树德、致新、树礼 6 个书院。

作为一所新创大学，学校根据国际化高水平研究型大学的办学定位和目标，面向国家和粤港澳大湾区战略新兴产业发展的重大需求，以理、工、医为主，兼具商科和特色人文社科的学科体系，快速建成本硕博人才培养体系。

学校目前已设置理学院、工学院等十大学院，建成 33 个院系及若干中心，开设 41 个本科专业。2018 年，南科大入选博士学位授权单位及硕士学位授权单位。建校至今，已获一级学科博士点 14 个，专业类别博士点 3 个，一级学科硕士点 15 个、专业类别硕士点 8 个，博士后流动站 7 个。

学校占地总面积 198 万平方米。目前规划总建筑面积为 171.62 万平方米，校园已建成总建筑面积 97.50 万平方米。校园二期项目工学院、理学院、商学院、第三教学楼等一批新建筑已全部投入使用。校园三期工程南科大附属医院（校本部）、医学院、教师宿舍等项目建设稳步推进，将陆续于 2026 年竣工交付。此外还孵化了深圳海洋大学，工程已开工建设，建筑面积 60.05 万平方米。校园环境优美、交通便利，校园内九山一水，风景如画。建筑彰显“厚重、节能、实用、环保”理念，成为绿色建筑和生态校园的典范，立志于建立中国最美的大学校园之一。

School of Science, SUSTech

Founded on December 6, 2017, the College of Science of Southern University of Science and Technology (SUSTech) is a basic and research-oriented college with a focus on basic sciences and related applied sciences and technologies. The College consists of five academic departments including Department of Mathematics, Department of Physics, Department of Chemistry, Department of Earth and Space Sciences, Department of Statistics and Data Science, and six research institutes: National Center for Applied Mathematics Shenzhen, SUSTech International Center for Mathematics, Shenzhen Grubbs Institute, Institute for Quantum Science and Engineering, Center for Advanced Light Sources Science, and Institute of Major Scientific Facilities for New Materials.

College of Science gathered and trained 21 academicians, 28 faculties have been awarded the National Science Fund for Distinguished Young Scholars, and 13 were nominated for the National Science Fund for Excellent Young Scholars. We have 196 tenuretrack professors, many of whom have been in Shenzhen Talent Peacock Plan, and Pearl River Talent Program and serve on the editorial board of international journals or have positions in overseas institutes.

College of Science aims at cultivating top-notch talents with a “solid foundation, inter-disciplinary competence, creativity and practice awareness”. With strength in basic disciplines, the College attaches great importance to students’ innovation spirit and practice capability. College of Science has 8 undergraduate programs, 4 postgraduate programs, 4 Ph.D. programs, and 4 post-doctoral research stations.

Based on the international platforms and the independent PI (Principal Investigator) system, experts at the College of Science have made great strides in their research, with an increasing number of papers being published in Science, Nature, and other renowned journals. The establishment of laboratories and the cooperation among the College, enterprises, and other research institutes have also thrived in the past few years.

With support from the Department of Science and Technology of Guangdong Province, the Development and Reform Commission of Shenzhen Municipality, and the Southern University of Science and Technology, we have established a series of major research platforms, including National Center for Applied Mathematics Shenzhen (NaCAMS), Guangdong Provincial Key Laboratory of Catalytic Chemistry, Guangdong Provincial Key Laboratory of Advanced Thermoelectric Materials and Device Physics, Shenzhen Key Laboratory of Deep Offshore Oil and Gas Exploration Technology, etc., creating favorable conditions for scientific breakthroughs.

By inheriting the spirit of exploration and innovation, the College of Science will continue to aim high and will contribute to the development of Guangdong - Hong Kong - Macau Greater Bay Area and the modernization of Shenzhen. Anchored in fundamental sciences, the College will vigorously expand its research scope to cutting-edge fields and strive for a world-class College of Science in the foreseeable future.

南方科技大学理学院

南方科技大学理学院成立于 2017 年 12 月 6 日，是以从事现代基础科学理论及其相关应用科学技术为主的基础性、研究型学院。学院现有数学系、物理系、化学系、地球与空间科学系、统计与数据科学系 5 个系，以及深圳国家应用数学中心、深圳国际数学中心（杰曼诺夫数学中心）、深圳格拉布斯研究院、量子科学与工程研究院、先进光源科学中心、新材料重大科技设施研究院 6 个研究机构。2022 年 2 月 14 日，南科大数学学科入选国家“双一流”建设学科名单，成为南科大和深圳市本土第一个入选学科。

学院现有教学科研序列教师（PI）197 位，其中，院士 18 位，国家杰青 / 青 A28 位，国家优青 / 青 B14 位，已初步建立起以院士为核心，以中青年学术骨干为主力的人才梯队。多人入选珠江学者、孔雀计划，担任国际学术期刊主编、副主编、编委，在国内外学术机构任职。

理学院坚持以培养“厚基础、宽口径、求创新、重实践”的复合型拔尖人才为目标，依托基础学科优势，注重学生创新精神与实践能力的培养。学院现已建成覆盖博士、硕士、学士完整教育层次的人才培养体系，设有 8 个本科专业、4 个硕士学位授权点、4 个博士学位授权点、4 个博士后科研流动站。学院现有在籍本科生 947 人，硕士研究生 873 人，博士研究生 712 人。

近年来，学院借助南科大国际化平台和独立 PI 制优势，创新创知氛围浓郁，在高水平成果突破、平台建设、产学研合作等方面取得了令人瞩目的成绩。学院现已获批 3 个国家级科研平台，12 个省级重点实验室 / 科研平台 / 中心，13 个市级科研平台 / 机构，12 个院级专业研究所或交叉研究中心以及 1 个校企合作实验室，多名专家教授在《自然》、《科学》等国际顶级学术期刊阐述重大研究成果，受邀发表评述文章。

依托国家战略性专项基金、广东省科技厅、深圳市发展改革委员会等部门以及南方科技大学的优质资源，我院正大力推动科研平台建设和产学研合作。目前已建立起深圳国家应用数学中心、广东省催化化学重点实验室、广东省可持续仿生材料与绿色能源重点实验室、广东省先进热点材料与器件物理重点实验室、深圳市深远海油气勘探技术重点实验室等大型科研平台。理学院科研设备先进齐全，科研团队实力雄厚，为科研创新、实验突破提供了有力支撑。

未来，理学院将继续秉持“探索世界，推动创新”的理念，面向粤港澳大湾区、珠三角以及深圳国际化现代化城市发展需求，围绕“教学、科研”两大主业推动学科交叉融合，不断拓展世界前沿科学研究新领域，与国内外优秀人才一道，争取在较短时间内把学院建设成国内外知名、世界一流的理学院。

Department of Statistics and Data Science, SUSTech

The Department of Statistics and Data Science at Southern University of Science and Technology (SUSTech) was established in April 2019, with the mission of building a world-class education and research base. The department aims to cultivate top-tier talents for the nation who possess a solid scientific foundation, active thinking, strong innovative awareness and competence, an international vision, a down-to-earth work ethic, and youthful ambition and ideals. It has established a complete talent training system covering bachelor's, master's, and doctoral programs. The department offers two undergraduate majors: Statistics, and Data Science and Big Data Technology. It is also authorized to confer master's and doctoral degrees in Probability Theory and Mathematical Statistics under the discipline of Mathematics. Its major research areas include mathematical statistics, biomedical statistics, financial statistics, and data science.

The department currently has 379 enrolled students, including 275 undergraduate students (majoring in Statistics and Data Science and Big Data Technology), 47 master's students, and 57 doctoral students. With the continuous development of the department and the expansion of its faculty, the number of graduate student admissions will increase year by year.

The faculty currently comprises 23 academic-track staff, including 3 Chair Professors, 4 Professors, 7 Associate Professors and 9 Assistant Professors, alongside 2 teaching-track faculty members.The department boasts an international and high-level faculty team, including: 1 Invited Speaker at the International Congress of Mathematicians (ICM), 2 recipients of the Second Class Prize of the State Natural Science Award, 1 Chang Jiang Chair Professor, 1 National Science Fund for Distinguished Young Scholars, 5 recipients of the National Overseas Talent Recruitment Program, 3 Fellows of the Institute of Mathematical Statistics (IMS), 2 Fellows of the American Statistical Association (ASA), 1 IMS Medallion Lecturer, 1 Fellow of the Royal Statistical Society (UK), 1 Fellow of the British Computer Society, 1 Guangdong Province Outstanding Young Scholar, 1 Guangdong Province Special Support Program for Young Top Innovative Talents, 1 Shenzhen Outstanding Talent Training Candidate, 1 Shenzhen High-Level National Leading Talent, 1 Shenzhen Excellent Young Scholar, and 2 Shenzhen Excellent Teachers. The Department of Statistics and Data Science is actively recruiting high-level talents from home and abroad.

The department fully leverages the strengths of the university and its school to gather high-quality teaching and research resources. It has established several research platforms, including the Big Data Innovation Center, the Biomedical Statistics Center, the Statistical Consulting Center, and the "Data + AI" Laboratory for Data Intelligence. In addition, the SUSTech Office of International Cooperation and the Department of Statistics and Data Science jointly provide high-level and diverse international education and development programs for undergraduate and graduate students, including summer and semester exchange programs at prestigious universities at home and abroad. These programs are designed to strengthen students' professional knowledge, broaden their social horizons and experience, and foster a diversified way of thinking.

南方科技大学统计与数据科学系

南方科技大学统计与数据科学系成立于 2019 年 4 月，以建设国际一流的教育培养和研究基地为目的。本系志在为国家培养出具有扎实的科学基础，思想活跃，创新意识和能力强，有国际视野，脚踏实地，有朝气、有理想的拔尖人才。本系已经建立起本硕博人才培养体系，拥有统计学和数据科学与大数据技术两个本科专业，及数学学科下概率论和数理统计硕博学位授予权，主要研究领域包括数理统计、生物医学统计、金融统计和数据科学。

本系现有在读本科生 275 人（含统计学专业、数据科学与大数据技术专业），在读硕士生 47 名、博士生 57 名。随着本系的发展和师资力量的增强，研究生招生数目将逐年上升。

目前共有 23 位教研序列教师，其中有讲席教授 3 人，教授 4 人，副教授 7 人，助理教授 9 人，并有教学序列教师 2 人。本系拥有国际化、高水平的师资队伍，包括 1 名国际数学家大会邀请报告人，2 名国家自然科学奖二等奖获得者，1 名长江讲席教授，1 名国家杰青，5 名国家海外引才计划，3 名国际数理统计学会 (IMS) 会士，2 名美国统计学会 (ASA) 会士，1 名 IMS Medallion 讲座演讲者，1 名英国皇家统计学会会士、1 名英国计算机学会会士，1 名广东省杰青，1 名广东省特支科技创新青年拔尖人才，1 名深圳市杰出人才培养对象，1 名深圳市高层次国家级领军人才，1 名深圳市优青和 2 名深圳市优秀教师。目前统计与数据科学系正在大力引进海内外高层次人才。

本系充分利用学校及学院的优势，汇聚优质教学和科研资源，目前已建立起大数据创新中心、生物医学统计中心、统计咨询中心和“Data+AI”数据智能实验室。同时，南科大国际合作部和统计与数据科学系为本科生及研究生提供高水平、多类型的“国际化教育培养拓展计划”，包括多个国内外知名大学的暑期或学期交换机会，旨在加强学生专业知识学习的基础上拓展学生的社会视野与社会阅历，培育学生多元化思维模式。

Faculty Position Announcement

Position Title: Tenure-track/Tenured faculty

Position Location: Shenzhen, Guangdong 518055, China

Position Description:

The Department of Statistics and Data Science invites applications and nominations for all ranks of tenure-track and tenured positions (Assistant Professor, Associate Professor, Full Professor and Chair Professor) in Statistics and Data Science. We seek applicants with outstanding methodological foundations in Statistics and Data Science. All areas of Statistics and Data Science will receive serious consideration with special emphasis on statistics, machine learning and AI, security and privacy, data mining and HCI. Candidates should have demonstrated excellence in research and a strong commitment to teaching. A doctoral degree is required at the time of appointment. Candidates for senior positions must have an established strong record of research and leadership in their research areas. The language of instruction can be either Chinese or English. Salary and start-up fund are highly competitive and will be commensurate with experience and qualifications. Regular faculty members (tenured and on tenure track) will be provided with housing subsidies. To apply, please send a cover letter, curriculum vitae, a research statement and a teaching statement to stat-dshire@sustech.edu.cn. Candidates should also arrange for at least three letters of recommendation sent directly to stat-dshire@sustech.edu.cn. Applications will be accepted until the positions are filled.

Supportive Conditions

SUSTech offers a globally competitive compensation and benefits package, along with a dynamic research environment. If selected for the Program, you will be appointed to a tenure-line position at SUSTech, with access to a wide array of supportive resources and favorable conditions, including but not limited to:

1）Compensation and Benefits

1. Compensation: Highly competitive salary with eligibility for personal income tax benefits in the Guangdong-Hong Kong-Macau Greater Bay Area.

2. Comprehensive benefits: Highest-standard Five Insurances and One Fund (social insurance and housing provident fund) in Shenzhen; supplementary commercial medical insurance and premium coverage for the Greater China region, along with a Shenzhen-standard annuity.

3. Personal grants: We will support you in applying for relevant talent programs aligned with national, provincial, and municipal talent initiatives. If selected, you may receive grants or other financial support.

2）Research Support

1. Research funding: Provide corresponding research funding support based on government policies and the outcome of your application. Special cases will be evaluated individually.

2. Team support: Research teams will be formed based on project needs. Postdoctoral researchers recruited within the approved quota will receive partial financial support from the university. You can also recruit additional postdocs using your own funds without limits. Graduate student quotas will be allocated based on university policies and departmental needs, and you have the flexibility to recruit research-line faculty and research/teaching assistants.

3. Office and Lab Facilities: High-quality office and laboratory spaces are available, alongside access to shared research platforms.

3）Living Benefits

1. Housing: Depending on availability, we offer either housing or housing subsidies.

2. Educational support for children: The SUSTech Education Group (Nanshan) includes 4 public schools in Nanshan District, 1 offering education from Grade 1 to 9, and 3 offering elementary education. For preschool, we have two public kindergartens: SUSTech Affiliated Kindergarten in the Nanshan District and a planned branch campus of the affiliated kindergarten located on our campus. We can assist with enrollment based on availability and government policies.

3. Medical Services: Access to healthcare through our 10 affiliated hospitals, 5 teaching hospitals, and 6 scientific research collaboration bases.

4. Living support: We assist with applications for Shenzhen hukou (household registration) or residence permits for you and your family, as well as the "You Yue Talent Card" (Guangdong provincial talent card) and "Shenzhen Talent Pass". These cards offer specialized services in housing, healthcare, education, and spouse employment in Guangdong and Shenzhen.

Application Materials Required:

To apply, please submit the following items directly to stat-dshire@sustech.edu.cn to complete your application:

- Cover Letter
- Curriculum Vitae
- Research Statement
- Teaching Statement
- Three or More Reference Letters

Further Info:

https://www.sustech.edu.cn/

https://stat-ds.sustech.edu.cn/

Ms. ZHOU <stat-dshire@sustech.edu.cn>

Department of Statistics and Data Science,

Southern University of Science and Technology,

1088 Xueyuan Boulevard,

Shenzhen, Guang Dong 518055

China

Logistics & Practical Information

Accommodation

The 2026 International Chinese Statistical Association (ICSA) China Conference offers a short-term course scheduled for June 26, with the main conference taking place from June 27 to June 29 at Convention Center, School of Business, Chi Wah Lecture Hall(Third Teaching Building), SUSTech. It is recommended that the participants may make room reservation on their own in light of their needs and the organizing committee will recommend the hotels that have signed the preferential price agreements. Nevertheless, there is no guarantee that the agreement price is necessarily the lowest. Please make comparison of the room rates on other platforms and use your discretion. Please refer to the table below for hotel information.

Special reminder: When booking and checking into hotels, please remind the service personnel that we are here to attend the 2026 ICSA China Conference to avoid missing the negotiation price. Please make reservations as early as possible, preferably by phone.

Hotel Information:

	Hotel name	No. of rooms	Contact person	Remark	House price
1	Genpla Hotel Shenzhen Nanshan	70 Double Rooms; 50 Twin Rooms	Reservation Phone: 18923459180 (Manager Fu)Email: reservation@genplahotel.com	0.8 kms away from the campus	Double Room: RMB 680 per room per night (with 1 breakfast);Twin Room: RMB 768 per room per night (with 2 breakfasts)
2	Vienna Good Sleep International Hotel (Tanglang Metro Station)	90 Double Rooms; 80 Twin Rooms	Reservation Phone: 19874134572 (Manager Zhou)Email: 15935000973@qq.com	1.5 kms away from the campus (Shuttle buses will be provided throughout the conference)	Double Room: RMB 448 per room per night (with 2 breakfasts);Twin Room: RMB 438 per room per night (with 2 breakfasts)
3	Shenzhen Nanshan Jingfeng Hotel (Xili University Town Branch)	280 Double Rooms; 30 Twin Rooms	Reservation Phone: 13728828561 (Manager Jin)Email: jh19981225@qq.com	4.5 kms away from the campus (Shuttle buses will be provided throughout the conference)	Double Room: RMB 388 per room per night (no breakfast); RMB 428 per room per night (with 1 breakfast); RMB 468 per room per night (with 2 breakfasts);Twin Room: RMB 428 per room per night (no breakfast); RMB 508 per room per night (with 2 breakfasts)
4	Sheraton Shenzhen Nanshan	70 Double Rooms; 80 Twin Rooms	Reservation Phone: 13923841909 (Manager Yang)Email: szxsn.rsvn@sheraton.com	3.6 kms away from the campus (Shuttle buses will be provided throughout the conference)	Double Room: RMB 1000 per room per night (with 1 breakfast);Twin Room: RMB 1100 per room per night (with 2 breakfasts)

住宿

2026 年泛华统计协会中国会议设有短期课程，短期课程时间为 6 月 26 日，主会议时间为 6 月 27 日 - 6 月 29 日，地点为南方科技大学会议中心、商学院、第三教学楼（智华教学楼）。建议参会人员根据需要自行预订酒店，会务组推荐协议酒店，但协议价格并不一定是最低价，请各位参会者对比其他平台的价格后作出最有利于自己的选择。酒店信息请参考下表。

特别提醒：参会人员预订和入住酒店时，请提示服务人员是来参加 2026 年泛华统计协会中国会议，以免错失协议价。请与会人员尽早预定，最好通过电话预定。

酒店信息：

	酒店名称	房间数量	联系人	备注	房 价
1	深圳南山深铁塘朗城君璞酒店	大床 70 间；双床 50 间	订房电话 :18923459180（付经理）；邮箱：reservation@genplahotel.com	距学校 0.8 公里	大床房 680 元 / 间 / 天 / 单早；双床房 768 元 / 间 / 天 / 双早
2	维也纳好眠国际酒店（塘朗地铁站店）	大床 90 间；双床 80 间	订房电话 :19874134572（周经理）；邮箱 :15935000973@qq.com	距学校 1.5 公里（会议期间往返有大巴接送）	大床房 448 元 / 间 / 天 / 双早；双床房 438 元 / 间 / 天 / 双早
3	深圳璟峯酒店（西丽大学城店）	大床 280 间；双床 30 间	订房电话 :13728828561（金经理）；邮箱 :jh19981225@qq.com	距学校 4.5 公里（会议期间往返有大巴接送）	大床房 388 元 / 间 / 天 / 无早餐；428 元 / 间 / 天 / 单早；468 元 / 间 / 天 / 双早；双床房 428 元 / 间 / 天 / 无早餐；508 元 / 间 / 天 / 双早
4	深圳南山博林天瑞喜来登酒店	大床 70 间；双床 80 间	订房电话 :13923841909（杨经理）；邮箱 :szxsn.rsvn@sheraton.com	距学校 3.6 公里（会议期间往返有大巴接送）	大床房 1000 元 / 间 / 天 / 单早；双床房 1100 元 / 间 / 天 / 双早



Getting to Southern University of Science and Technology at Shenzhen

1. Via Shenzhen Bao'an International Airport

a. Online Ride-hailing / Taxi: Directly to Southern University of Science and Technology, with a driving time of approximately 40-50 minutes and an estimated cost of 80-120 RMB (You can use online ride-hailing apps such as Didi Chuxing and Amap, which support English operation);

Online Ride-hailing Pick-up Area A (Outdoor) is located outside Gate 15 of the Ground Transportation Center. Ride-hailing Process: Set the pick-up location as "Shenzhen Airport Online Ride-hailing Area A". In Area A, connect with drivers and vehicles according to lane colors and pick-up location signs, and board the vehicle in order to leave.

Online Ride-hailing Pick-up Area B (Indoor) is located on the 1st floor of the Ground Transportation Center (GTC) (you



can take the elevator down from the atrium of the Ground Transportation Center to reach it). Ride-hailing Process: Set the pick-up location as "Shenzhen Airport Online Ride-hailing Area B (Reserved Welcome Area)". After entering the waiting area in Area B, enter the license plate number on the electronic screen to apply for vehicle access authorization, and board the vehicle after it arrives at the pick-up location.



The taxi pick-up point is located outside Gate 13 of the airport's Ground Transportation Center.



b. Subway Travel: Take the escalator or elevator from Terminal T3 or the Ground Transportation Center to the basement floor (B1) to take the subway at Airport Station.



Specific Route: Take Metro Line 11 (Airport Line) to Qianhaiwan Station, transfer to Metro Line 5 (Loop Line) to Tanglang Station, exit from Exit C and walk about 5 minutes (350 meters) to reach Gate 1 of the campus. The whole journey takes about 1 hour, and the fare is 7 RMB.

Key Reminder: Since the last train of Line 5 departs at 23:00, you must arrive at Qianhaiwan Station before 23:00 to make the transfer. It is recommended that you take the Line 11 train before 22:45 at Airport Station to reserve sufficient transfer time.

Shenzhen Metro has upgraded its international convenient services. You can bind an overseas bank card to WeChat to activate the Shenzhen Metro Mini Program QR Code, use the overseas version of Digital RMB for payment, or purchase tickets at intelligent POS machines supporting 6 mainstream foreign cards such as VISA and Master Card at hub stations like Airport Station. Airport Station is also equipped with multilingual intelligent customer service terminals (supporting more than 10 languages including English, Japanese and Korean), providing real-time translation and consultation services throughout the process.

c. Airport Bus: Purchase tickets at the Airport Express Service Counter next to Gate 17 on the 2nd floor of the airport or through the "Shenzhen Airport Bus Ticketing" Mini Program. You can take the official airport bus Line NA2 (Airport — North Station West Square Emergency Parking Lot) currently in operation at Shenzhen Airport. The fare is 25 RMB, operating from 0:00 to 4:00 every day, with a departure every 40 minutes (no buses from 2:41 to 3:59). Take it to North Station, then transfer to an online ride-hailing or taxi to Southern University of Science and Technology.



d. Bus: The pick-up point is located outside Gate 16 of the Ground Transportation Center. You can take Bus M527 to Lixiang Park, then transfer to Bus 81 at the same station to Southern University of Science and Technology.

2. Via Shenzhen North Station

a. Direct Online Ride-hailing / Taxi to Southern University of Science and Technology, with a driving time of approximately 15 minutes and an estimated cost of 20-35 RMB;

The online ride-hailing pick-up areas at Shenzhen North Station are divided into Area A and Area B: Area A is located on the B1 floor of the West Square, and Area B is located on the 1st floor of the West Square.





There are three taxi pick-up points at Shenzhen North Station: the 1st floor platform of the East Square, the 2nd floor platform of the East Square, and inside the Passenger Station of the West Square. Exit from Exit A1 or A2 of the East Square, walk about 50 meters along the edge of the square or take the escalator/staircase to the 2nd floor platform, then walk about 30 meters to reach the taxi waiting area and board the taxi in line; or exit from Exit B1 or B2 of the West Square, turn right and walk about 100 meters along the main road to enter Shenzhen North Station Passenger Station, and the taxi waiting area is located on the north side of the station.

b. Subway: Exit from Exit A1 of the East Square, take the elevator in front down to take Metro Line 5 (Loop Line) to Tanglang Station, exit from Exit C and walk about 5 minutes (350 meters) to reach Gate 1 of Southern University of Science and Technology. The journey takes about 25 minutes, and the fare is 3 RMB;

c. Bus: Exit from Exit A1 of the East Square, take the elevator in front down to take the bus, take Bus M217/81 to Southern University of Science and Technology Station, then walk about 80 meters to arrive. The journey takes about 30 minutes, and the fare is 2 RMB.

3. Via Futian Station

After exiting the station, take Metro Line 3 to Shaonian Gong Station. Transfer to Line 4 at the same station and get off at Shenzhen North Station. Then, transfer to Line 5 and alight at Tanglang Station. Exit the station and walk to Southern University of Science and Technology. The whole journey takes about 50 minutes and costs 5 yuan. If taking a taxi or ride-hailing service, the trip takes approximately 30 minutes.

4. Via Shenzhen Railway Station

After exiting the station, walk to Luohu Metro Station. Take Metro Line 1 to Convention & Exhibition Center Station. Transfer to Line 4 at the same station and get off at Shenzhen North Sta-

tion. Then, transfer to Line 5 and alight at Tanglang Station. Exit the station and walk to Southern University of Science and Technology. The entire journey takes about 60 minutes and costs 6 yuan. If taking a taxi or ride-hailing service, the trip takes around 40 minutes.

5. Via Hong Kong

Take the high-speed train from West Kowloon Station in Hong Kong to Futian Station or Shenzhen North Station. Then, transfer to Metro Line 5 and alight at Tanglang Station. Exit the station and walk to Southern University of Science and Technology. The whole journey takes about 60 minutes. If taking a taxi or cross-border ride-hailing service, the trip takes approximately 50 minutes.

Supplementary Notes

1. Currently, Gates 1, 3 and 6 of Southern University of Science and Technology are open for vehicles. Online ride-hailing and taxis can directly reach the school gate. It is recommended to set the location as "Southern University of Science and Technology Gate 1" (main entrance), and foreign guests can directly show the location to the driver;

2. The operation hours of buses and subways are subject to the actual schedule on the day. It is recommended to reserve sufficient travel time during peak hours (7:30-9:00, 17:30-19:00);

3. All transportation fares and driving times are reference values. The actual cost and time may fluctuate slightly due to road conditions, vehicle type and exchange rate. You can consult the corresponding transportation service hotline in advance for confirmation;

4. Foreign guests need to prepare valid documents such as passports and visas in advance, cooperate with border inspection when entering Shenzhen, and it is recommended to learn about China's entry-related regulations in advance.

前往南方科技大学

1. 经深圳宝安国际机场

a. 网约车 / 出租车：直达南方科技大学，车程约 40-50 分钟，预计 80-120 元（可使用滴滴打车、高德地图等网约车 APP，支持英文操作）；

网约车乘车点 A 区（室外）位于地面交通中心 15 号门外，打车流程：打车点位设为“深圳机场网约车 A 区”，A 区按车道颜色、上客位标识对接司机和车辆，依序上车离开。



网约车乘车点 B 区（室内）位于地面交通中心（GTC）一层（地面交通中心中庭乘坐电梯下楼即可到达）。打车流程：打车点位设为“深圳机场网约车 B 区（预约迎客区）”，B 区进入候车区后，在电子屏幕处输入车牌预约车辆通行授权，车辆到达上客点位后上车离开。



出租车乘车点位于机场地面交通中心 13 号门外。



b. 地铁出行：从机场 T3 航站楼或地面交通中心乘扶梯、直梯至负一层机场站乘坐地铁。



具体路线：乘坐地铁 11 号线（机场线）至前海湾站，换乘 5 号线（环中线）至塘朗站，从 C 口出站后步行约 5 分钟（350 米）即可抵达校区 1 号门，全程约 1 小时，票价 7 元。

关键提醒：由于 5 号线末班车为 23:00，您必须在 **23:00 前到达前海湾站**才能换乘。建议您在机场站乘坐 **22:45 之前**的 11 号线列车，预留充足的换乘时间。

深圳地铁已实现国际化便捷服务升级，可通过微信绑定境外银行卡开通深圳地铁小程序乘车码、使用数字人民币境外版支付，或在机场站等枢纽站点通过支持 VISA、Master Card 等 6 种主流外卡的智能 POS 机购票，机场站更设有多语种智能客服终端（支持英、日、韩等 10 余种语言），全程提供实时翻译咨询服务。

c. 机场大巴：在机场二楼 17 号门旁机场快线服务柜台或通过“深圳机场巴士售票”小程序购票，可乘坐深圳机场当前运营的正规机场大巴 NA2 线（机场—北站西广场应急停车场），票价 25 元，每日 0:00-4:00 运行，每 40 分钟一班（2:41-3:59 无车），乘坐至北站后，再换乘网约车或出租车至南方科技大学。



d. 公交：乘车点位于地面交通中心 16 号门外，可乘坐 M527 路到荔香公园后，同站换乘 81 路车到南方科技大学。

2. 经深圳北站

a. 网约车 / 出租车直达南方科技大学，车程约 15 分钟，预计 20-35 元；

深圳北站的网约车上客区分为A区和B区: A区位于西广场B1层，B 区位于西广场 1F 层。





深圳北站的出租车上车点共有三处：东广场一层平台、东广场二层平台、西广场客运站内。东广场 A1、A2 出口出站，沿广场边缘走约 50 米或乘坐扶梯 / 步梯上二层平台，走约 30 米到达出租车候车区，排队上车；或从西广场 B1、B2 出口出站，右转沿主干道走约 100 米，进入深圳北站汽车客运站，站内北侧即为出租车候客区。

b. 地铁：东广场 A1 出口出来，前面电梯下去乘坐地铁 5 号线（环中线）至塘朗站，从 C 口出站后步行约 5 分钟（350 米）抵达南方科技大学 1 号门，用时约 25 分钟，票价 3 元；

c. 公交：东广场 A1 出口出来，前面电梯下去乘坐公交，乘公交车（M217/81 路）至南方科技大学站，步行约 80 米抵达，用时约 30 分钟，票价 2 元。

3. 经福田站

出站乘坐地铁 3 号线至少年宫地铁站，站内换乘 4 号线至深圳北站地铁站，站内换乘 5 号线至塘朗地铁站，出站步行至南方科技大学，全程约 50 分钟，票价 5 元。

出租车或网约车，全程约 30 分钟。

4. 经深圳站（火车站）

出站步行至罗湖地铁站，乘坐地铁 1 号线至会展中心地铁站，站内换乘 4 号线至深圳北站地铁站，站内换乘 5 号线至塘朗地铁站，出站步行至南方科技大学，全程约 60 分钟，票价 6 元。

出租车或网约车，全程约 40 分钟。

5. 经香港

从香港西九龙站乘坐高铁至福田站或深圳北站，换乘地铁 5 号线至塘朗地铁站，出站步行至南方科技大学，全程约 60 分钟。

出租车或跨境网约车，全程约 50 分钟。

补充说明

1. 南方科技大学现有 1 号门、3 号门、6 号门通车，网约车 / 出租车可直达校门口，建议定位“南方科技大学 1 号门”（主入口），国外客人可直接出示定位给司机；

2. 公交、地铁运营时间以当日实际为准，高峰期（7:30-9:00、17:30-19:00）建议预留充足出行时间；

3. 各类交通票价、车程为参考值，实际费用及耗时可能因路况、车型、汇率略有波动，可提前咨询对应交通服务热线确认；

4. 国外客人需提前准备好护照、签证等有效证件，入境深圳时配合边防检查，建议提前了解中国入境相关规定。

Medical Information

1. Emergency Phone

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

24-Hour Alarm and Help Hotline of the University Security Office: 0755-88010110, 18566211101

24-Hour Emergency Contact Number of Campus Medical Service Center:18218715551

Emergency Contact Person of the Organizing Committee: Lan LI: 0755-88015665, 15013768385

2. Recommended Hospitals

(1) Southern University of Science and Technology Community Health Service Center

Services: Basic diagnosis and treatment, simple trauma treatment, health consultation, first-aid equipment available, and routine inspections can be carried out.

Address: 1st Floor, Student Services Building, Southern University of Science and Technology (adjacent to the Student Activity Center)

Consultation Hotlines: 0755-88010499, 0755-88010120

(2) Shenzhen University General Hospital

Address: No. 1098 Xueyuan Avenue, Taoyuan Street, Nanshan District, Shenzhen

Advantages: A Grade A tertiary public hospital with comprehensive medical disciplines and the shortest driving distance from SUSTech.

Consultation Hotline: 0755-21839999

(3) Southern University of Science and Technology Hospital

Address: No. 6019 Liuxian Avenue, Xili, Nanshan District, Shenzhen

Advantages: A tertiary comprehensive public hospital.

Consultation Hotlines: 0755-25232188-3030 (Medical Consultation), 0755-26999580 (Feedback)

(4) The University of Hong Kong-Shenzhen Hospital

Address: No. 1 Haiyuan First Road, Futian District, Shenzhen

Advantages: A tertiary grade A hospital adopting Hong Kong-style management, with a comprehensive international medical center, rich experience in foreign-related medical services and smooth English communication.

Consultation Hotlines: 0755-86913333 (Chinese), 0755-86913366 (English), 0755-86913388 (International Medical Center)

3. Pharmacies

(1) Zhengguan Pharmaceutical Pharmacy

Address: No. 119, Building 1, Keda Yayuan, No. 1099 Xueyuan Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen

(2) Nanbei Pharmacy (Keda Yayuan Store)

Address: No. 116A, Building 1, Keda Yayuan, No. 1099 Xueyuan Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen

(3) Huatai Pharmaceutical Pharmacy (Baoneng Global Center Xili Store)

Address: B1 Floor, Baoneng Global Center, No. 3288 Liuxian Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen

(4) Huatai Pharmaceutical Pharmacy (Tanglang City Branch)

Address: No. 109, Commerce of Block A, Block B and Block C, Tanglang City Plaza (West District), No. 3333 Liuxian Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen

(5) Neptune Star Health Pharmacy (Baoneng Global Center Store)

Address: No. 103, Basement 1, Block B, Baoneng Global Center, No. 3288 Liuxian Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen

It is recommended to carry enough personal prescription drugs for the entire trip, and keep the original packaging and doctor's prescription. Familiarize yourself with China's regulations on imported drugs (especially drugs containing anesthetic and psychotropic ingredients) to avoid carrying prohibited items.

1. **紧急电话**

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

学校保卫部 24 小时报警求助电话：0755-88010110、18566211101

学校社康 24 小时紧急联系电话：18218715551

组委会紧急联系人：李兰，0755-88015665，15013768385

2. **推荐医院**

（1）南方科技大学社区健康服务中心

服务：基础诊疗、简单外伤处理、健康咨询，配备急救设备，可开展常规检验。

地址：南方科技大学学生服务楼一楼（紧邻学生活动中心）

咨询电话：0755-88010499、0755-88010120

（2）深圳大学总医院

地址：深圳市南山区桃源街道学苑大道 1098 号

优势：三级公立医院，学科齐全，距离南科大车程最近。

咨询电话：0755-21839999

（3）南方科技大学医院

地址：深圳市南山区西丽留仙大道 6019 号

优势：三级综合性公立医院。

咨询电话：0755-25232188-3030（就医咨询）、0755-26999580（意见反馈）

（4）香港大学深圳医院

地址：深圳市福田区海园一路 1 号

优势：三级甲等，香港式管理，国际医疗中心完善，涉外医疗经验丰富，英文沟通顺畅。

咨询电话：0755-86913333（中文）、0755-86913366（英语）、0755-86913388（国际医疗）

3. **药店**

（1）正观医药大药房

地址：深圳市南山区桃源街道福光社区学苑大道 1099 号科大雅苑 1 栋 119

（2）南北药行 (科大雅苑店)

地址：深圳市南山区桃源街道福光社区学苑大道 1099 号科大雅苑 1 栋 116A

（3）华太医大药房 (宝能环球汇西丽店)

地址：深圳市南山区桃源街道福光社区留仙大道 3288 号宝能环球汇 B1

（4）海王星辰健康药房（宝能环球汇店）

地址：深圳市南山区桃源街道福光社区留仙大道 3288 号宝能环球汇 B 馆负一楼 103

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

Weather Notice

1. Weather overview

(1)Temperature

Daytime temperatures are typically between 30°C-35°C (86°F-95°F), while nighttime temperatures generally range from 26°C-30°C (79°F-86°F). The high humidity may make it feel even hotter.

(2)Precipitation

During the conference period (summer), Shenzhen experiences hot and rainy weather, with frequent short but heavy thunderstorms, often occurring in the afternoon or evening. There is also a possibility of prolonged heavy rainfall.

(3)Sunshine & Humidity

Intervals between rainfall are often accompanied by intense sunshine and high UV radiation, with generally high humidity levels (averaging around 80%). The combination of heat and humidity can be particularly noticeable.

Suggestion: Please check real-time weather updates from the Shenzhen Meteorological Bureau frequently and plan your travel accordingly. During peak heat hours, minimize outdoor activities. If symptoms of heatstroke such as dizziness, fatigue, or nausea occur, move to a shaded area immediately, hydrate, and seek assistance if needed.

2. Typhoon response

Shenzhen's summer is a typhoon-prone season. Please pay close attention to typhoon warning signals issued by meteorological authorities (Blue → Yellow → Orange → Red, with Red being the highest level).

Follow instructions from the conference organizers and local emergency authorities.

During typhoon impacts, avoid unnecessary outings and stay indoors away from windows and doors.

Familiarize yourself with emergency evacuation routes at the conference venue and your accommodation in advance.

Suggestion: Starting one week before the conference, stay updated via the "Shenzhen Weather" official app, the China Meteorological Administration website, or the Hong Kong Observatory for the latest forecasts and typhoon tracking information.

3. Dressing and Equipment Suggestions

Clothing: Lightweight, quick-dry, and breathable short-sleeved tops are recommended. Bring a long-sleeved layer or sun-protective clothing for strong air-conditioning indoors and sun protection outdoors.

Shoes and socks: Comfortable, breathable walking shoes are advisable. Waterproof or quick-dry shoes may be useful for rainy days. Bring extra pairs of cotton socks.

Rain gear: A wind-resistant folding umbrella or a lightweight raincoat is recommended (rain in Shenzhen is often accompanied by gusts).

Sun protection:High-SPF sunscreen (SPF50+ PA+++), sunglasses, and a wide-brimmed hat or a dual-use sun/rain umbrella.

Mosquito repellent: Mosquitoes are common in Shenzhen during summer. Use repellents containing DEET or Picaridin for better effectiveness.

Others: Portable fan, water bottle, tissues/wet wipes, power bank, and common heatstroke prevention medicines (e.g., Huoxiang Zhengqi Shui).

天气须知

1. 天气概况

(1) 温度

日间温度通常在 30° C-35° C (86° F-95° F) 之间，夜间温度通常在 26° C-30° C (79° F-86° F) 之间。体感温度因高湿度可能更高。

(2) 降水

会议期间 (夏季) 为深圳高温多雨季节，多突发性短时雷阵雨，常见于午后或傍晚，偶有持续性强降雨。降水概率较高，出行需做好防雨准备。

(3) 日照与湿度

降雨间隙常伴随强烈日照与高紫外线指数，空气湿度普遍较高（平均约 80%）。闷热感明显，需注意防暑。

建议：建议随时关注深圳气象局发布的实时天气预报，合理规划出行时间。高温时段尽量减少户外停留，如出现头晕、乏力、恶心等中暑症状，请立即至阴凉处休息、补水并及时求助。

2. 台风应对

深圳夏季为台风多发期，请密切关注气象部门发布的台风预警信号（蓝色→黄色 →橙色 → 红色，红色为最高级别）。

遵循会议组织方与当地应急部门的指引。

台风影响期间尽量避免外出，留在坚固的室内，远离门窗。

提前了解会议地点及住宿场所的应急疏散路线。

建议：会前一周起，可通过“深圳天气”官方 APP、中国气象局官网、香港天文台等渠道持续关注天气动态与台风路径预报。

3. 着装与装备建议

衣物：以轻薄、速干、透气的短袖衣物为主，可备长袖薄外套或防晒衣（用于室内强冷空调环境及户外防晒）。

鞋袜：建议穿着舒适、透气的步行鞋，可备一双防水或速干鞋应对雨天。多备几双棉质袜子。

雨具：推荐携带抗风折叠伞或轻便雨衣（深圳降雨常伴随阵风）。

防晒：高倍数防晒霜（SPF50+ PA+++）、太阳镜、遮阳帽或晴雨两用伞。

防蚊虫：深圳夏季蚊虫较多，建议使用含避蚊胺（DEET）或派卡瑞丁（Picaridin）成分的驱蚊产品。

其他：便携小风扇、水壶、纸巾及湿巾、充电宝、常用防暑药品（如藿香正气水等）。

Local Attractions

Welcome to Shenzhen, a dynamic city that combines modern vitality with profound cultural heritage. Below are carefully selected attractions with detailed introductions, suitable for relaxing during breaks from conferences or exploring the city's unique charm.

1. Wutong Mountain National Park

Wutong Mountain National Park is Shenzhen's largest urban natural oasis, just a short drive from the city center. Covered with dense subtropical forests, it boasts fresh mountain air, quiet shaded hiking trails and a serene atmosphere, making it the perfect escape from busy conference schedules. Hiking to its 943.7-meter main peak (Shenzhen's highest point), you can enjoy a breathtaking panoramic view of the entire Shenzhen skyline, overlooking skyscrapers, the Shenzhen Bay and surrounding mountain ranges. The park is completely free to visit, with multiple easy hiking routes suitable for casual short walks. It's the best place to relax, recharge and feel the balance of urban vitality and natural peace.





2. Lianhua Mountain Park

Located in the center of the city, it is a lush urban oasis with dense vegetation and a serene atmosphere. Climbing to the top of the mountain, you can overlook the entire urban skyline, feel the vitality of the city, and enjoy a moment of peace away from the busy conference schedule. It is free to visit and perfect for a short walk.



3.Shenzhen Bay Park

As a coastal ecological and leisure park, it stretches along the coast with an open view and a quiet environment. You can take a walk, ride a bicycle, or simply sit and enjoy the sea breeze. It is free to visit and easily accessible by public transport, making it an ideal place to relax and escape the hustle and bustle of conferences.



4. Talent Park

A coastal park with the theme of “talent culture”, integrating natural scenery and modern design. The environment is beautiful, the trails are comfortable, and it is a good place to take a walk, relax and feel the innovative vitality of the city.



5. Dameisha Beach Park

A famous coastal scenic spot with wide and clean beaches, clear sea water and soft sand. You can walk along the beach, enjoy the sea breeze, and experience the unique coastal style of Shenzhen, which is a perfect choice for relaxing and enjoying the sea view.



6. Overseas Chinese Town Ecological Square

It has an elegant environment and excellent greening, surrounded by artistic facilities such as art galleries and cultural centers. It is a comfortable leisure space with a strong artistic atmosphere, suitable for walking and relaxing.



7. Shenzhen Science and Technology Museum (New Venue, Guangming)

Shenzhen Science and Technology Museum (New Venue) in Guangming District is the world's first science museum themed on "digital civilization and innovative future". Designed by Zaha Hadid Architects, its futuristic spaceship-like exterior has become a viral city landmark. With 128,300 square meters of exhibition space, it offers fully immersive interactive experiences: from cutting-edge communication technology displays, hands-on scientific experiments to futuristic digital art installations. Most exhibition areas are free to visit (online reservation required in advance). It's perfect for exploring global tech innovation, experiencing the charm of digital civilization, and having a fun, inspiring short visit.



8. Shenzhen International Art Museum (SIMoA)

Located in the core of Guangming Science City, Shenzhen International Art Museum (SIMoA) is a brand-new world-class cultural landmark designed by Chinese Academy of Engineering academician Cui Kai. With the design concept of "standing together in solidarity", its sleek modern architecture blends perfectly with the surrounding lush Science Park. As a comprehensive international art hub, it hosts top-tier contemporary art exhibitions, and features professional exhibition halls, a dedicated children's art gallery, open-air theater, and multi-functional art studios. Most public exhibitions are free to enter with advance reservation. It's an ideal spot to immerse in global art, capture stunning architectural photos, and enjoy a quiet cultural break from your busy conference schedule.



9.Shenzhen Museum

A comprehensive museum that systematically displays the city's development history, folk customs and marine culture. The exhibits are rich and professional, allowing you to deeply understand the development context of Shenzhen and gain a comprehensive understanding of the city's culture and history.



10. Ping An Finance Center Observation Deck

As one of the landmark high-rise buildings in Shenzhen, its observation deck offers a 360-degree panoramic view of the city. You can overlook the bustling urban scenery, feel the modern charm of the city, and it is equipped with complete facilities to provide a comfortable visiting experience. It is a great place to understand the modern style of Shenzhen.



11. Nantou Ancient City

With a history of nearly 1,700 years, it is a precious historical relic in Shenzhen. The ancient streets, traditional buildings and modern cultural and creative spaces coexist, allowing you to experience the integration of history and modernity, and feel the profound cultural heritage of the city.



12. Dapeng Fortress

A well-preserved coastal defense site from the Ming and Qing dynasties, with complete ancient city patterns and traditional Lingnan architectural styles. It is a quiet and peaceful place, suitable for a slow tour to understand the historical context of Shenzhen and feel the charm of traditional culture.

24-Hour Food & Beverage Establishments

1.Rongshanhe Mutton Noodles (Baoneng Store)

No. 126, Semi-basement, Building 3, West Zone, Baoneng City Garden, Tanglang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen

Tel:19867707761

2.Jin Zhaopai Fresh-cut Noodles & Casserole Soup

Shop 15-02, Building 15, Chongwen Garden, No. 3888 Liuxian Avenue, Fuguang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen

Tel:15766200168

3.Chuqu Shuicheng Lamb Noodles (South University of Science and Technology Branch)

No. 110, Building 1, Ke Daya Yuan, No. 1099 Xueyuan Avenue, Fuguang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen City

Tel: 18938882672

4.Xinrongji Guilin Rice Noodles (Xili Branch)

No. 18-3 Xinguang Road, Wenyuan Building, Xili Community, Xili Subdistrict, Nanshan District, Shenzhen City

Tel: 18377467708

5.He Xi Dian Afternoon Tea · Cakes & Bakery (Xili Branch)

No. 112, Building 1, Liyuan 2nd Village, No. 33-2 Wenyuan Street, Xili Community, Xili Subdistrict, Nanshan District, Shenzhen City

Tel: 13691769885

6.New York Burger – Beef Burger (Shenzhen University Town Branch)

Tangxing Road, Building W5, Tanglang Industrial Zone A, Tanglang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen City (Address: No. 162 Tangxing Road)

Tel: 17727434895

7.Wallace · Whole Chicken Burger (Nanshan Tanglang Branch)

No. 109, Qidian Apartment, No. 107 Xiangrui 1st Road, Tanglang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen City

Tel: 18820156406

8.KFC Delivery (Xili Baoneng City Branch)

Podium Building, Block B & C, East Zone, Baoneng City Garden, No. 3288 Liuxian Avenue, Fuguang Community, Taoyuan Subdistrict, Nanshan District, Shenzhen City

Tel: 021-59097718

Tips:

The 24-hour food & beverage establishments listed in this handbook are statistically collected with valid information up to May 2026, aiming to provide convenient dining references for participants arriving late at night in Shenzhen. The service quality (including but not limited to taste, hygiene and food freshness) and on-site operating conditions of these venues may be adjusted without prior notice due to temporary closure, renovation, updated business hours or other reasons, and we are unable to conduct real-time verification. We kindly suggest confirming the latest opening status and service information through mapping applications or phone calls before your visit. The organizing team will not be liable for any inconvenience caused by personal taste differences, temporary venue adjustments or other unexpected situations.

www.icsa.org



INTERNATIONAL CHINESE STATISTICAL ASSOCIATION
泛華統計協會



南方科技大學
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY