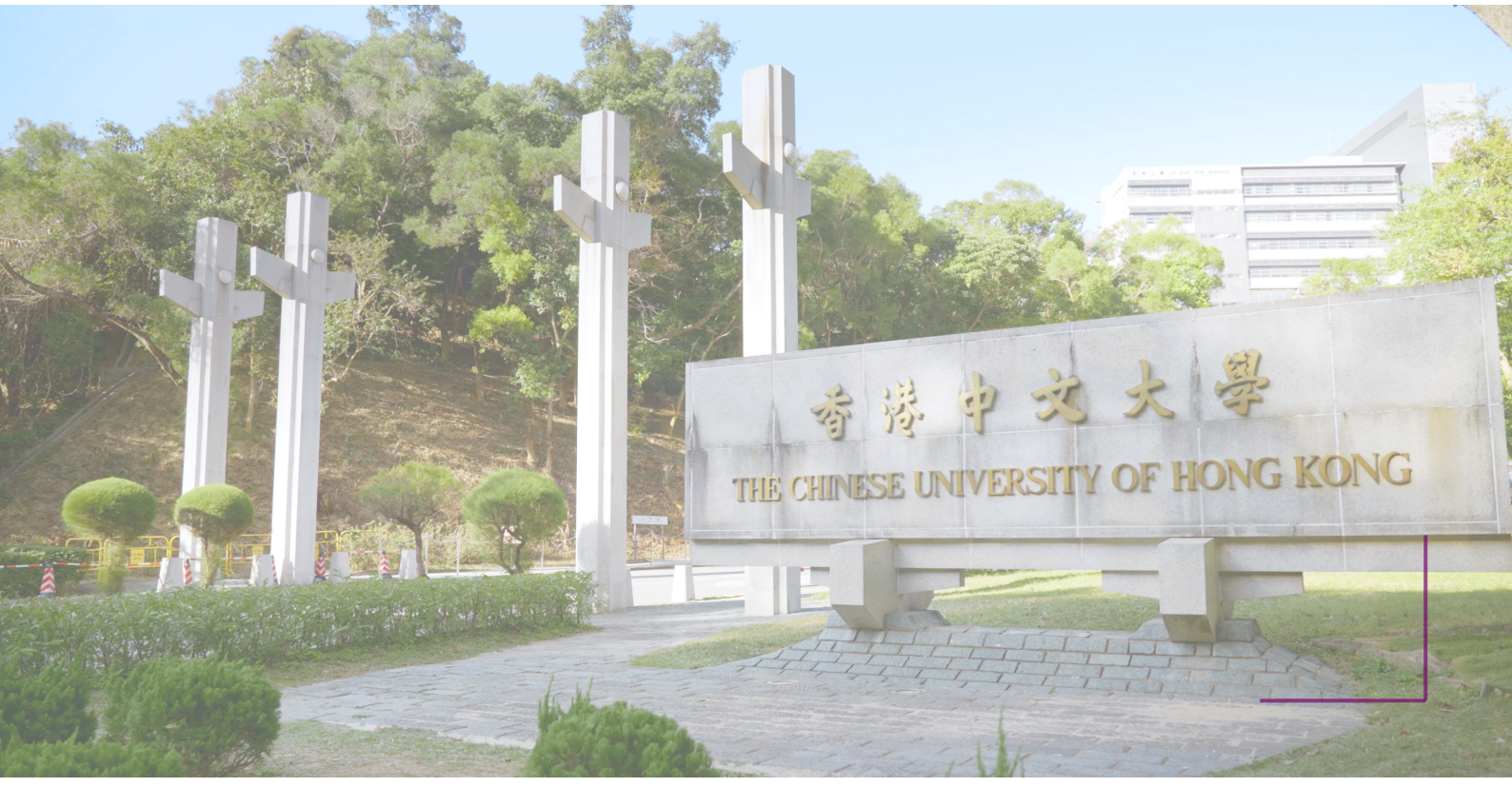
ims Asia Pacific Rim Meeting 2026

The Seventh IMS Asia Pacific Rim Meeting

June 13 – 16, 2026 • Hong Kong



Welcome to IMS-APRM 2026

Dear Colleagues and Friends,

It is our great pleasure to welcome you to the **Seventh Institute of Mathematical Statistics Asia Pacific Rim Meeting (IMS-APRM 2026)**, held in Hong Kong from **June 13 to June 16, 2026**.

The IMS-APRM series brings together researchers from the Asia-Pacific region and around the world to present and discuss recent developments in statistics, probability, and their applications. Since its inception, the IMS-APRM has grown into one of the premier international conferences in the mathematical sciences, fostering collaboration and exchange of ideas across diverse research communities.

This year's meeting features an outstanding scientific program, including:

- **2 Plenary Lectures** by Professor Andrea Montanari (Stanford University) and Professor Hans-Georg Müller (UC Davis)
- **21 Distinguished Lecture sessions** covering cutting-edge topics in statistics and probability
- **72 Invited Paper sessions** organized by leading researchers from around the world
- **14 Contributed sessions** showcasing emerging research
- **2 Poster sessions** for interactive discussions

The conference is hosted at the **Henry Cheng International Conference Centre** in the Cheng Yu Tung Building at The Chinese University of Hong Kong (CUHK). Nestled in the lush hills of the New Territories, CUHK offers a beautiful and inspiring setting for academic exchange.

We extend our sincere gratitude to the Scientific Program Committee, the Local Organizing Committee, all session organizers, speakers, and participants for making this meeting possible. We also thank our sponsors — the Institute of Mathematical Statistics, the Bernoulli Society, and our regional and institutional partners — for their generous support.

We hope you will enjoy the scientific program, the social events, and the vibrant city of Hong Kong. We wish you a productive and memorable conference experience.

Byeong U. Park
Seoul National University
Co-chair, Scientific Program Committee

Xiaotong T. Shen
University of Minnesota
Co-chair, Scientific Program Committee

Junhui Wang
The Chinese University of Hong Kong
Co-chair, Local Organizing Committee

Xinyuan Song
The Chinese University of Hong Kong
Co-chair, Local Organizing Committee

Committees

Scientific Program Committee

Co-chairs:

- **Byeong U. Park**, Seoul National University, South Korea
- **Xiaotong T. Shen**, University of Minnesota, United States

Members:

- Peter Radchenko, The University of Sydney
- Parthanil Roy, Indian Institute of Technology Bombay
- Tomonari Sei, The University of Tokyo
- Chae Young Lim, Seoul National University
- Fang Yao, Peking University
- Ming-Yen Cheng, Hong Kong Baptist University
- Yingcun Xia, National University of Singapore
- Hsin-Cheng Huang, Academia Sinica
- Junhui Wang, The Chinese University of Hong Kong
- Ajay Jasra, The Chinese University of Hong Kong, Shenzhen
- Ying Zhang, University of Nebraska Medical Center
- Soutir Bandyopadhyay, Colorado School of Mines

Local Organizing Committee

Co-chairs:

- **Junhui Wang**, The Chinese University of Hong Kong
- **Xinyuan Song**, The Chinese University of Hong Kong

Members:

All members are from The Chinese University of Hong Kong unless otherwise noted.

- Chun Man Chan
- Kin Wai Chan
- Ben Dai
- Xiaodan Fan
- Xiao Fang
- Isaac Sze Him Leung
- Gen Li
- Zhixiang Lin
- Yuanyuan Lin
- Yanny Yan Yau Ng
- Ming Ouyang
- Wendy Sze Wan Tang
- Tony Sit
- Tiejun Tong *(Hong Kong Baptist University)*
- Alan Tze-Kin Wan *(CityU of Hong Kong)*
- Yingying Wei
- Hoi Ying Wong
- Julian Ying Wang Wong
- Phillip Yam
- Can Yang *(HKUST)*
- Dan Yang *(The University of Hong Kong)*
- Chun Yip Yau
- Jiacheng Zhang
- Xingqiu Zhao *(Hong Kong Polytechnic University)*
- Huichen Zhu

General Information

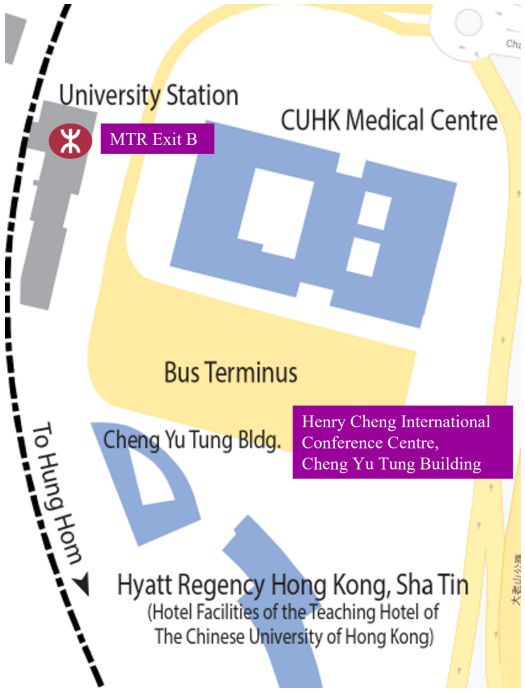
Venue

The conference is held at the **Henry Cheng International Conference Centre**, located in the **Cheng Yu Tung Building** at The Chinese University of Hong Kong (CUHK).

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

Getting There:

- **MTR:** Take the East Rail Line to *University Station*. The Cheng Yu Tung Building is directly accessible from the station Exit B (less than 5 minutes walk).
- **Taxi:** Ask for "Cheng Yu Tung Building, CUHK" (鄭裕彤樓, 香港中文大學).



Map: Cheng Yu Tung Building, CUHK (University Station, East Rail Line)

Registration Desk

The registration desk is located in the **lobby of the Cheng Yu Tung Building** (1/F).

Hours:

- June 13 (Saturday): 08:00 – 17:00
- June 14 (Sunday): 08:00 – 17:00
- June 15 (Monday): 08:00 – 12:00
- June 16 (Tuesday): 08:00 – 15:00

Wi-Fi

Complimentary Wi-Fi is available throughout the conference venue. Connection details will be provided at the registration desk.

Meals & Coffee Breaks

- **Coffee breaks** will be served in the tea reception areas on both the 1st and 2nd floors during the scheduled break times.
- **Lunch boxes** will be provided in the restaurant – The Stage on the 3rd and 5th floors from 11:30 until 14:00.

Session Format

The scientific program includes three types of sessions:

- **Invited Paper (IP) Sessions:** Each session features four speakers, each presenting for **25 minutes** (including Q&A).
- **Distinguished Lecture (DL) Sessions:** Each session features one Distinguished Lecturer presenting for **50 minutes**, followed by two invited speakers each presenting for **25 minutes** (including Q&A).
- **Contributed Session (CS) Sessions:** Each session features multiple speakers, each presenting for **20 minutes** (including Q&A).

Social Events

- **Poster Session 1:** June 13 (Saturday), 17:20 – 18:30
- **Conference Banquet:** June 14 (Sunday), 19:00. Venue details will be announced at the conference.
- **Half-day Break:** June 15 (Monday) afternoon – free time for sightseeing, networking, or individual appointments.
- **Poster Session 2:** June 16 (Tuesday), 15:30 – 16:30

Emergency Contacts

- **Emergency Services (Police/Fire/Ambulance):** 999
- **CUHK Campus Security:** +852 3943 7999
- **Conference Secretariat:** Details available at the registration desk

Plenary Speakers

Plenary Lecture 1 — **June 13 (Saturday), 09:30 – 10:30** — **Room: LT1A**

Professor Andrea Montanari

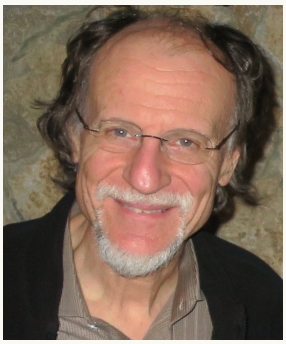
Stanford University
Departments of Statistics and Mathematics

Andrea Montanari is a Professor of Statistics and Mathematics at Stanford University. He is a leading researcher in high-dimensional statistics, machine learning, and information theory. His work spans a broad range of topics including random matrix theory, graphical models, optimization in high dimensions, and the statistical physics of inference problems.

Professor Montanari is a Fellow of the Institute of Mathematical Statistics, a Fellow of the IEEE, and has received numerous awards including the Le Cam Lecture Prize. His research has had a profound impact on our understanding of computational and statistical trade-offs in modern data science.

Plenary Lecture 2 — **June 14 (Sunday), 09:00 – 10:00** — **Room: LT1A**

Professor Hans-Georg Müller



University of California, Davis
Department of Statistics

Hans-Georg Müller is a Distinguished Professor of Statistics at the University of California, Davis. He is one of the world's foremost experts in functional data analysis, nonparametric statistics, and the analysis of longitudinal data. His pioneering contributions to functional data analysis have shaped the field and opened new directions in the study of complex data objects.

Professor Müller is a Fellow of the Institute of Mathematical Statistics, a Fellow of the American Statistical Association, and an elected member of the International Statistical Institute. He has received the Humboldt Research Award and has served on numerous editorial boards of leading statistics journals.

At-a-Glance Schedule

DL Distinguished Lecture IP Invited Paper CS Contributed PS Poster Plenary Plenary / Special

Time	LT1A	LT1B	209A	209B	203	201	202	214	215
June 13 (Saturday)									
09:00–09:30	Welcome and Opening Ceremony (IMS Committee) (LT1A)								
09:30–10:30	Plenary Talk 1 (LT1A)								
10:30–10:50	*Tea Break*								
10:50–12:30	DL02	DL12	IP06	IP09	IP08	IP61	IP25	IP45	IP47
12:30–13:30	*Lunch Break*								
13:30–15:10	DL05	DL01	IP04	IP70	IP24	IP44	IP60	IP59	IP69
15:10–15:30	*Tea Break*								
15:30–17:10	DL03	DL10	IP16	IP27	IP30	IP26	IP33	IP37	
June 14 (Sunday)									
09:00–10:00	Plenary Talk 2 (LT1A)								
10:00–10:20	*Tea Break*								
10:20–12:00	DL18	DL11	IP40	IP03	IP10	IP17	IP05	IP11	IP23
12:00–13:00	*Lunch Break*								
13:00–14:40	DL09	DL16	IP14	IP18	IP28	IP38	IP36	IP07	CS07
14:50–16:30	DL06	DL04	IP01	IP19	IP22	IP41	IP35	IP02	CS04
16:30–16:50	*Tea Break*								
16:50–18:30	DL21	IP62	IP48	IP50	PS	IP46	IP13	CS10	CS09
19:00	Conference Banquet								
June 15 (Monday)									
09:00–10:40	DL07	DL08	IP56	IP15	IP58	IP34	IP54	IP55	CS03
10:40–11:00	*Tea Break*								
11:00–12:40	DL17	DL20	IP63	IP67	IP72	IP39	IP66	CS01	CS13
Afternoon	Half-day Break								
June 16 (Tuesday)									
09:00–10:40	DL13	DL14	IP29	IP12	IP21	IP52	IP49	IP32	CS02
10:40–11:00	*Tea Break*								
11:00–12:40	DL15	DL19	IP53	IP64	IP68	IP71	CS05	CS06	CS11
12:40–13:30	*Lunch Break*								
13:30–15:10	IP43	IP31	IP57	IP20	IP51	IP42	CS08	CS12	CS14

Detailed Program

DL Distinguished Lecture IP Invited Paper CS Contributed Plenary Plenary/Special

June 13 (Saturday)

09:00–09:30

09:00–09:30 **Welcome and Opening Ceremony (IMS Committee)**

09:30–10:30

Plenary Talk 1 —Professor Andrea Montanari Room: LT1A

10:50–12:30

DL02 — Causal Inference Room: LT1A

June 13 (Saturday) 10:50–12:30 Abstracts

Chair: Zhenhua Lin, National University of Singapore, Singapore

Distinguished Lecturer: Jane-Ling Wang, University of California, Davis, United States
"Quantile Treatment Effect Estimation from Censored Lifetime Data"

- **Qixian Zhong**, Xiamen University, China
  "Generative Doubly Robust Estimation for General Treatment Effects"
- **Peng Ding**, University of California, Berkeley, United States
  "Estimating treatment effects with competing intercurrent events in randomized controlled trials"

DL12 — Hierarchical Poisson Species Sampling Models Room: LT1B

June 13 (Saturday) 10:50–12:30 Abstracts

Chair: Ming-Yen Cheng, Hong Kong Baptist University, Hong Kong

Distinguished Lecturer: Lancelot James, The Hong Kong University of Science and Technology, Hong Kong
"Poisson Hierarchical Indian Buffet Processes –with indications for Microbiome Models and other possibilities"

- **Yi Li**, University of Michigan, United States (Discussant)
  "Discussant"
- **Edwin Fong**, The University of Hong Kong, Hong Kong
  "Predictive inference for grouped data"

IP06 — Recent Advances in Causal Inference Room: 209A

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Xinran Li, The University of Chicago, United States
Chair: Will Wei Sun, Purdue University, United States

- **Ben Hansen**, University of Michigan, United States
  "Design-based Hájek estimation after allocation by cluster and within blocks"
- **Yuhao Wang**, Tsinghua University, China
  "Asymptotic theory of the best-choice rerandomization using the Mahalanobis distance"
- **Zhichao Jiang**, Sun Yat-sen University, China
  "Principled analysis of crossover designs: causal effects, efficient estimation, and robust inference"
- **Xinran Li**, The University of Chicago, United States
  "Randomization Inference with Sample Attrition"

IP09 — Recent Developments on Generative Models

Room: 209B

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Chuanhai Liu, Purdue University, United States
Chair: Xiao Wang, Purdue University, United States

- **Jian Huang**, The Hong Kong Polytechnic University, Hong Kong
  "TBD"
- **Xiaotong Shen**, University of Minnesota, United States
  "Manifold-Aligned Generative Transport"
- **Xiaoyue Niu**, The Pennsylvania State University, United States
  "A multilayer network model for aggregated relational data"

IP08 — Recent Developments on Network Analysis and Biostatistics

Room: 203

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Xiaotong Shen, University of Minnesota, United States
Chair: Jie Ding, University of Minnesota, United States

- **Ji Zhu**, University of Michigan, United States
  "Statistical Inference for Latent Space Models of Network Data with Edge Covariates"
- **Haoran Xue**, City University of Hong Kong, Hong Kong
  "MR2G: A novel framework for causal network inference using GWAS summary data"
- **Lu Tian**, Stanford University, United States
  "An Honest Cross-Validation Estimator for Prediction Performance"
- **Annie Qu**, University of California, Irvine, United States
  "Dynamic Topic Modeling with a Higher-Order Hypergraphical Representation"

IP61 — Advances in the Design and Analysis of Clinical Trials

Room: 201

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Yang Liu, Renmin University of China, China
Chair: Yang Liu, Renmin University of China, China

- **Wei Zhang**, Chinese Academy of Sciences, China
  "Optimal treatment allocations accounting for population differences"
- **Yu Jun**, Beijing Institute of Technology, China
  "Minimum free energy driven randomized allocation to improve covariate balance"
- **Yang Liu**, Renmin University of China, China
  "The Impact of Unobserved Covariates on Covariate-Adaptive Randomized Experiments"
- **Sai Li**, Tsinghua University, China
  "Personalizing black-box models for nonparametric regression with minimax optimality"

IP25 — Advances in Statistical Methods for Complex Data and Clinical Studies

Room: 202

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Le Zhou, Hong Kong Baptist University, Hong Kong
Chair: Le Zhou, Hong Kong Baptist University, Hong Kong

- **Lijian Yang**, Tsinghua University, China
  "Simultaneous inference for finite distribution functions of stochastic processes"
- **Lucy Xia**, The Hong Kong University of Science and Technology, Hong Kong
  "Statistical Inference with Mixed-Effect Model for Covariate-Adaptive Randomized Experiments"
- **Shunan Yao**, Hong Kong Baptist University, Hong Kong
  "U-processes and their application in mean estimation"
- **Yifan Chen**, Hong Kong Baptist University, Hong Kong
  "A Recipe for Causal Graph Regression: Confounding Effects Revisited"

IP45 — Applied Probability and Financial Mathematics

Room: 214

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Li-Hsien Sun, National Central University, Taiwan
Chair: Li-Hsien Sun, National Central University, Taiwan

- **Ju-Yi Yen**, University of Cincinnati, United States
  "An arbitrage driven price dynamics of Automated Market Makers in the presence of fees"
- **Chi Seng Pun**, Nanyang Technological University, Singapore
  "Pairs Trading with Frictions: Transaction Costs, Market Impact, and Optimal Strategies"
- **Chuan-Hsiang Han**, National Tsing Hua University, Taiwan
  "Accelerating Deep Learning by Efficient Importance Sampling for CVaR Estimation"
- **Li-Hsien Sun**, National Central University, Taiwan
  "Partial Information in a Mean-Variance Portfolio Selection Game"

IP47 — Design and Analysis of Modern Experiments

Room: 215

June 13 (Saturday) 10:50–12:30 Abstracts

Organizer: Frederick Kin Hing Phoa, Academia Sinica, Taiwan
Chair: Frederick Kin Hing Phoa, Academia Sinica, Taiwan

- **Ming-Chung Chang**, Academia Sinica, Taiwan
  "Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs"
- **Jing-Wen Huang**, Academia Sinica, Taiwan
  "Cyclic Order-of-addition experiments"
- **William Li**, Shanghai Advanced Institute of Finance, China
  "Robust Integer-Valued Designs for Non-Linear Models: An Algorithmic Approach with Efficiency Guarantee"
- **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan

13:30–15:10

DL05 — To be confirmed

Room: LT1A

June 13 (Saturday) 13:30–15:10 Abstracts

Distinguished Lecturer: Jing Lei, Carnegie Mellon University, United States
"Evaluating Black-Box Classifiers via Stable Adaptive Two-Sample Inference"

- **David Rügamer**, University of Munich, Germany
  "Smooth Optimization for Sparse Learning via Overparameterization"

DL01 — Structured Nonparametrics

Room: LT1B

June 13 (Saturday) 13:30–15:10 Abstracts

Chair: Byeong U. Park, Seoul National University, South Korea
Distinguished Lecturer: Enno Mammen, Heidelberg University, Germany

"Additive nonparametric regression: interaction, high dimension and beyond"

- **Wolfgang Polonik**, University of California, Davis, United States
  "On the structure of the Euler characteristic of a VR-filtration"
- **Changwon Choi**, Seoul National University, South Korea
  "Additive Fréchet regression for random objects"

IP04 — Statistical Opportunities in Deep and Reinforcement Learning

Room: 209A

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Hongtu Zhu, The University of North Carolina at Chapel Hill, United States / Guohao Shen, The Hong Kong Polytechnic University, Hong Kong
Chair: Alexandra Carpentier, University of Potsdam, Germany

- **Guohao Shen**, The Hong Kong Polytechnic University, Hong Kong
  "Symmetries in Deep Neural Networks and Implications to Learning"
- **Faming Liang**, Purdue University, United States
  "Uncertainty Quantification for Physics-Informed Neural Networks with Extended Fiducial Inference"
- **Chengchun Shi**, The London School of Economics and Political Science, United Kingdom
  "Demystifying LLM Reasoning through the Lens of U-Statistics"
- **Hongtu Zhu**, The University of North Carolina at Chapel Hill, United States
  "Causal deepsets for off-policy evaluation under spatial or spatio-temporal interferences"

IP70 — Recent Development on AI and Biostatistics

Room: 209B

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Xiaotong Shen, University of Minnesota, United States
Chair: Annie Qu, University of California, Irvine, United States

- **Jie Ding**, University of Minnesota, United States
  "The Future of AI Scientists: Emerging Directions and Fundamental Challenges"
- **Ji Yuan**, The University of Chicago, United States
  "From research idea to peer-reviewed draft: an agentic pipeline for statistical methodology research with adversarial revision"
- **Joan J. Ren**, University of Maryland, United States
  "Empirical Likelihood Based Multivariate Distribution Estimator for Various Types of Censored Data and Its Application in Censored Causal Inference"

IP24 — Classification, Community Detection and Inference for High Dimensional Complex Data

Room: 203

June 13 (Saturday) 13:30–15:10 Abstracts

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

- **Jinchi Lv**, University of Southern California, United States
  "HNCI: High-Dimensional Network Causal Inference"
- **Yin Xia**, Fudan University, China
  "A Unified Framework for Large-Scale Inference of Classification: Error Rate Control and Optimality"
- **Yi-Hsin Yang**, National Health Research Institutes, Taiwan
  "Determining lines of therapy algorithms to detect cancer progression events in electronic health records"

IP44 — Recent Advances in Sampling Methods

Room: 201

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Alexandre Thiery, National University of Singapore, Singapore
Chair: Alexandre Thiery, National University of Singapore, Singapore

- **Zi Yang Meng**, The University of Hong Kong, Hong Kong

"(Artificial) intelligent Monte Carlo sampling in quantum many-body systems"

- **Zijing Ou**, Imperial College London, United Kingdom
  "Neural Flow Samplers: Improved Training and Architectures"
- **Miha Bresar**, The Chinese University of Hong Kong-Shenzhen, China
  "TBD"
- **Jiajun He**, University of Cambridge, United Kingdom
  "Diffusion Model Control with Monte Carlo Methods"

IP60 — Some of the Latest Advances in Multiple Testing

Room: 202

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Dennis Leung, The University of Melbourne, Australia
Chair: Dennis Leung, The University of Melbourne, Australia

- **Jesse Hemerik**, Erasmus University Rotterdam, Netherlands
  "Resampling-based multi-resolution false discovery exceedance control"
- **Shinjini Nandi**, Montana State University, United States
  "Leveraging the group structure of hypotheses for more powerful multiple testing with FDR control for the filtered rejection set"
- **Jinzhou Li**, National University of Singapore, Singapore
  "Simultaneous false discovery proportion bounds via knockoffs and closed testing"
- **Qiuqi Wang**, Georgia State University, United States
  "False discovery rates of refreshing monitoring procedures"

IP59 — Advances in Forecasting and Time Series Analysis

Room: 214

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Anastasios Panagiotelis, Monash University, Australia
Chair: Michele Guindani, University of California, Los Angeles, United States

- **Yu Fu**, Melbourne Business School, Australia
  "Regression Copula Process MIDAS for Macroeconomic Density Forecasting"
- **Chao Wang**, The University of Sydney, Australia
  "Combining Forecasts of Value-at-Risk and Expected Shortfall Forecasts When Many Methods Are Available"
- **Xiaoqian Wang**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
  "Online conformal inference for multi-step time series forecasting"
- **Shanika Wickramasuriya**, Monash University, Australia
  "Hierarchical forecasting: The role of information"

IP69 — Recent Developments on AI and Statistics

Room: 215

June 13 (Saturday) 13:30–15:10 Abstracts

Organizer: Xiaotong Shen, University of Minnesota, United States
Chair: Yixuan Qiu, Shanghai University of Finance and Economics, China

- **Xuan Bi**, University of Minnesota, United States
  "Distribution-Preserving Watermarking for Mixed-Typed Data"
- **Helen Zhang**, University of Arizona, United States
  "Dynamic Supervised Principal Component Analysis for Classification"
- **Jian Shi**, Chinese Academy of Sciences, China
  "TBD"
- **Lei Li**, Chinese Academy of Sciences, China
  "Empirical Lossless Compression Bound of a Data Sequence"

15:30–17:10

DL03 — Semiparametric Modeling and Predictive Inference

Room: LT1A

June 13 (Saturday) 15:30–17:10 Abstracts

Chair: Yu Gu, The University of Hong Kong, Hong Kong

Distinguished Lecturer: Huixia Judy Wang, Rice University, United States
"Semiparametric Distribution Learning and Predictive Inference: A Quantile Regression Process Perspective"

- **Tony Sit**, The Chinese University of Hong Kong, Hong Kong
  "Censored Quantile Regression with Time-Dependent Covariates and its Extensions"
- **Seyoung Park**, Yonsei University, South Korea
  "Smoothed Quantile Additive Regression with Functional Lasso Kernel Regularization in High Dimensions"

DL10 — Advances in Nonparametric Statistics and Functional Data Analysis

Room: LT1B

June 13 (Saturday) 15:30–17:10 Abstracts

Chair: Peter Radchenko, The University of Sydney, Australia

Distinguished Lecturer: Aurore Delaigle, The University of Melbourne, Australia
"Nonparametric estimators of nonstationary densities of streaming data"

- **Frédéric Ferraty**, Toulouse Mathematics Institute, France
  "Nonparametric Approach to Forecasting Density-Valued Time Series"
- **Dominik Liebl**, University of Bonn, Germany
  "Making Event Study Plots Honest: A Functional Data Approach to Causal Inference"

IP16 — Recent Advances in Network Analysis and Dependence Testing for Complex Data

Room: 209A

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Qin Fang, The University of Sydney, Australia
Chair: Wanjie Wang, National University of Singapore, Singapore

- **Fan Wang**, The University of Melbourne, Australia
  "Smooth and abrupt changes in autoregressive tensor models."
- **Zihan Wang**, Tsinghua University, China
  "Time Series Gaussian Chain Graph Models"
- **Qing Jiang**, Beijing Normal University, China
  "Statistical inference for high-dimensional logistic regression with latent random change point"
- **Qin Fang**, The University of Sydney, Australia
  "Large-Scale Multiple Testing of Cross-Covariance Functions with Applications to Functional Network Models."

IP27 — Extremes, Heavy Tails and Stable Processes

Room: 209B

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Ayan Bhattacharya, Indian Institute of Technology Bombay, India
Chair: Junho Yang, Academia Sinica, Taiwan

- **Zaoli Chen**, University of Science and Technology of China, China
  "Extremal Phase Transitions under Long-Range Dependence"
- **Bikramjit Das**, Singapore University of Technology and Design, Singapore
  "Measuring extreme tail association"
- **Ayan Bhattacharya**, Indian Institute of Technology Bombay, India
  "Asymptotic behavior of extremes of regularly varying branching random walk on time-inhomogeneous random tree"

IP30 — Statistical Methods for High-dimensional and Complex Data Structures Room: 203

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Anirvan Chakraborty, Indian Institute of Science Education and Research, Kolkata, India
Chair: Yei Eun Shin, Seoul National University, South Korea

- **Kento Egashira**, Tokyo University of Science, Japan
  "Change-point detection for mean and covariance structures in high-dimensional data under a strongly spiked eigenvalue model"
- **Anne van Delft**, Columbia University, United States
  "Analyzing shape in a time series of random geometric objects"
- **Han Lin Shang**, Macquarie University, Australia
  "Conformal prediction for high-dimensional functional time series"
- **Anirvan Chakraborty**, Indian Institute of Science Education and Research, Kolkata, India
  "Near-perfect Clustering Based on Recursive Binary Splitting Using Max-MMD"

IP26 — Advances in High-dimensional Statistical Inference and Dynamic Modeling Room: 201

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Lucy Xia, The Hong Kong University of Science and Technology, Hong Kong
Chair: Lucy Xia, The Hong Kong University of Science and Technology, Hong Kong

- **Shu-Chin Lin**, National Taiwan University, Taiwan
  "Domain Selection for Functional Linear Models"
- **Weichen Wang**, The University of Hong Kong, Hong Kong
  "On reference-regulated multiperiod mean-variance portfolio optimization in high dimensions"
- **Lijia Wang**, City University of Hong Kong, Hong Kong
  "Network autoregression for binary responses in social networks"
- **Le Zhou**, Hong Kong Baptist University, Hong Kong
  "Adaptive Testing and Estimation for High-Dimensional Covariance Change-Points"

IP33 — Recent Advances in Copula Theory and Applications Room: 202

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Xiaoling Dou, International Christian University, Japan
Chair: Xiaoling Dou, International Christian University, Japan

- **Alexander McNeil**, York University, United Kingdom
  "Measures and Models of Non-Monotonic Dependence"
- **Issey Sukeda**, The University of Tokyo, Japan
  "Dependence modeling for circular data in neuroscience"
- **Jia-Han Shih**, National Sun Yat-sen University, Taiwan
  "A class of regression association measures based on concordance"
- **Martial Longla**, University of Mississippi, United States
  "On estimation problems based on new types of copulas"

IP37 — Additive Models and High-dimensional Inference: Modern Tools for Complex Data Room: 214

June 13 (Saturday) 15:30–17:10 Abstracts

Organizer: Young Kyung Lee, Kangwon National University, South Korea
Chair: Young Kyung Lee, Kangwon National University, South Korea

- **Jeong Min Jeon**, Seoul National University, South Korea
  "Hilbertian additive regression with general estimated variables."
- **Eun Ryung Lee**, Sungkyunkwan University, South Korea
  "Parallel Additive Regression for High-Dimensional Data: Efficient Computation with Convergence and Consistency Guarantees"

- **Seong J. Yang**, Jeonbuk National University, South Korea
  "A regularization approach for time-dependent AUC in survival analysis using deep neural networks"
- **Ming-Yen Cheng**, Hong Kong Baptist University, Hong Kong
  "Sparse optimal model averaging under a general framework"

June 14 (Sunday)

09:00–10:00

Plenary Talk 2 —Professor Hans-Georg Müller — Room: LT1A

10:20–12:00

DL18 — Advancing Precision Health with AI and Statistics — Room: LT1A

June 14 (Sunday) 10:20–12:00 Abstracts

Chair: Yingying Wei, The Chinese University of Hong Kong, Hong Kong

Distinguished Lecturer: Hsin-Chou Yang, Academia Sinica, Taiwan
"Unlocking Precision Health: Insights from Genetics, Medical Imaging, and Multimodal AI Integration"

- **Jung-Ying Tzeng**, North Carolina State University, United States
  "Statistical learning for polygenic risk prediction and CNV association testing"
- **Tso-Jung Yen**, Academia Sinica, Taiwan
  "An Automatic Approach to Explainable AI with Applications to Medical Image Classification"

DL11 — Statistical Inference in Metric Spaces — Room: LT1B

June 14 (Sunday) 10:20–12:00 Abstracts

Chair: Fang Yao, Peking University, China

Distinguished Lecturer: Xueqin Wang, University of Science and Technology of China, China
"Metric Distribution Function: A New Statistical Cornerstone for Non-Euclidean Data"

- **Ting Li**, Southern University of Science and Technology, China
  "Ball Impurity: Measuring Heterogeneity in General Metric Spaces"
- **Jin Zhu**, University of Birmingham, United Kingdom
  "Identification of Genetic Factors Associated with Corpus Callosum Morphology: Conditional Strong Independence Screening for Non-Euclidean Responses"

IP40 — Advances in Statistical Machine Learning — Room: 209A

June 14 (Sunday) 10:20–12:00 Abstracts

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

- **Ben Dai**, The Chinese University of Hong Kong, Hong Kong
  "EnsLoss: Stochastic Calibrated Loss Ensembles for Preventing Overfitting in Classification"
- **Hui Zou**, University of Minnesota, United States
  "Double Descent in the Enhanced Response Envelope Model"
- **Yufeng Liu**, University of Michigan, United States
  "Low-rank Reinforcement Learning with Heterogeneous Human Feedback"
- **Yoonkyung Lee**, The Ohio State University, United States
  "Data Influence Dynamics under Iterative Training Algorithms"

IP03 — Advances in Graph Learning and Network Analysis

Room: 209B

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Hao Chen, University of California, Davis, United States
Chair: Ji Zhu, University of Michigan, United States

- **Jie Peng**, University of California, Davis, United States
  "Inferring Latent Graphs from Stationary Signals"
- **Wanjie Wang**, National University of Singapore, Singapore
  "Node Differential Privacy in Node Ranking for Social Networks"
- **Xin Tong**, The University of Hong Kong
  "Stance Drift in AI-mediated communication"
- **Hao Chen**, University of California, Davis, United States
  "Community detection across mixing patterns for two or more communities"

IP10 — Statistics and AI

Room: 203

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Yongdai Kim, Seoul National University, South Korea
Chair: Yongdai Kim, Seoul National University, South Korea

- **Ilsang Ohn**, Inha University, South Korea
  "Adaptive online Bayesian inference via expert aggregation"
- **Masaaki Imaizumi**, The University of Tokyo, Japan
  "High-dimensional theory for dynamics of neural network training and transformer inference"
- **Kiseop Lee**, Purdue University, United States
  "Attention based reading, highlighting, and forecasting of the limit order book"

IP17 — Statistical Learning Theory with Dependent/Markov Observations

Room: 201

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Pierre Alquier, ESSEC Business School, Singapore
Chair: Pierre Alquier, ESSEC Business School, Singapore

- **Azadeh Khaleghi**, ENSAE Paris, France
  "On the Estimation of Mixing Coefficients from Stationary Ergodic Sample Paths"
- **Geoffrey Wolfer**, Waseda University, Japan
  "Optimistic Estimation of Convergence in Markov Chains with the Average-Mixing Time"
- **Vahe Karagulyan**, ESSEC Business School, France
  "Empirical PAC-Bayes bounds for Markov chains"
- **Daniel Paulin**, Nanyang Technological University, Singapore
  "Scalable MCMC methods for Bayesian learning of time series models"

IP05 — Distribution Shift and Data Integration: State of the Art and Future Outlook

Room: 202

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Ingrid Van Keilegom, KU Leuven, Belgium / Jiwei Zhao, University of Wisconsin–Madison, United States
Chair: Jiwei Zhao, University of Wisconsin–Madison, United States

- **Jiwei Zhao**, University of Wisconsin–Madison, United States
  "Towards the Efficient Inference by Incorporating Automated Computational Phenotypes under Covariate Shift"
- **Seong-ho Lee**, University of Seoul, South Korea
  "Semiparametric Framework for Efficient Semi-supervised Inference under Label Shift"
- **Molei Liu**, Peking University, China
  "Transfer Learning of CATE with Kernel Ridge Regression"
- **Chi-Shian Dai**, National Cheng Kung University, Taiwan

"Multiclass Classification Utilizing Heterogeneous External Machine Learning Predictions"

IP11 — Statistical and Algorithmic Foundation of Diffusion Models

Room: 214

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Yuting Wei, University of Pennsylvania, United States
Chair: Yuting Wei, University of Pennsylvania, United States

- **Yuxin Chen**, University of Pennsylvania, United States
  "Towards a Unified Framework for Guided Diffusion models"
- **Andre Wibisono**, Yale University, United States
  "Optimal Score Estimation via Empirical Bayes Smoothing"
- **Yuejie Chi**, Yale University, United States
  "Polynomial Convergence of Riemannian Diffusion Models"
- **Arnak Dalalyan**, ENSAE/CREST, France
  "Discretisation error of Denoising Diffusions measured in Wasserstein Distance"

IP23 — Advances in Causal Inference and Statistical Testing

Room: 215

June 14 (Sunday) 10:20–12:00 Abstracts

Organizer: Ying Yang, Fudan University, China
Chair: Ying Yang, Fudan University, China

- **Yifan Cui**, Zhejiang University, China
  "Double Machine Learning of Continuous Treatment Effects with General Instrumental Variables"
- **Guoyu Zhang**, Peking University, China
  "Bidirectional causal inference: a nonparametric potential outcome framework"
- **Zhengtian Zhu**, Tongji University, China
- **Xiaoyu Hu**, Xian Jiaotong University, China
  "Neural Wasserstein Two-Sample Tests"

13:00–14:40

DL09 — To be confirmed

Room: LT1A

June 14 (Sunday) 13:00–14:40 Abstracts

Distinguished Lecturer: Zhiliang Ying, Columbia University, United States
"TBD"

- **Yunxiao Chen**, The London School of Economics and Political Science, United Kingdom
  "TBD"
- **Jing Ouyang**, The University of Hong Kong, Hong Kong
  "Statistical Analysis of Large-scale Item Response Data under Measurement Non-invariance: A Statistically Consistent Method and Its Application to PISA 2022"

DL16 — Nonstandard Asymptotics in Causal Mediation Analysis

Room: LT1B

June 14 (Sunday) 13:00–14:40 Abstracts

Chair: Junhui Wang, The Chinese University of Hong Kong, Hong Kong
Distinguished Lecturer: Ian McKeague, Columbia University, United States
"On accurately calibrating test statistics at singularities in multidimensional parameter spaces"

- **Moulinath Banerjee**, University of Michigan, United States
  "'HARMLESS' sampling for determining level sets of a response function"
- **Guosheng Yin**, The University of Hong Kong, Hong Kong
  "Bayesian Knockoff Filter for Controlling False Discovery Rate"

IP14 — Recent Developments in Statistical Machine Learning and Causal Inference Room: 209A

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Yufeng Liu, University of Michigan, United States
Chair: Doudou Zhou, National University of Singapore, Singapore

- **Quefeng Li**, The University of North Carolina at Chapel Hill, United States
  "Inference on the Significance of Modalities in Multimodal Generalized Linear Models"
- **Zhengling Qi**, The George Washington University, United States
  "InSPO: Unlocking Implicit Self-Reflection for LLM Preference Optimization"
- **Will Wei Sun**, Purdue University, United States
  "When Is an LLM Really Better? Uncertainty-Aware Evaluation from Sparse Human Preferences"
- **Gongjun Xu**, University of Michigan, United States
  "Identifiability and Inference for Generalized Latent Factor Models"

IP18 — Inference under Heterogeneity and Side Information Room: 209B

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Gourab Mukherjee, University of Southern California, United States
Chair: Eun Ryung Lee, Sungkyunkwan University, South Korea

- **Kabir Verchand**, University of Southern California, United States
  "Statistical-computational gaps in estimation with missing not at random data"
- **Oh-Ran Kwon**, The Ohio State University, United States
  "Black-Box Knowledge Transfer for Distinct Feature Sets"
- **Fei Xue**, Purdue University, United States
  "An Empirical Bayes Regression for Multi-tissue Gene Expression Prediction"
- **Keisuke Yano**, The Institute of Statistical Mathematics, Japan
  "Inference under Frequency-Domain Heterogeneity via Spectral Rényi Divergences"

IP28 — Memory and Learning in Probability Room: 203

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Moumanti Podder, Indian Institute of Science Education and Research, Pune, India
Chair: Chi Seng Pun, Nanyang Technological University, Singapore

- **Shuhei Mano**, The Institute of Statistical Mathematics, Japan
  "Symmetric quantum walks on Hamming graphs and their limit distributions"
- **Anish Sarkar**, Indian Statistical Institute, New Delhi, India
  "Scaling limit of a drainage network model on perturbed lattice"
- **Neeraja Sahasrabudhe**, Indian Institute of Science Education and Research, Mohali, India
  "Elephant Random Walk with Tampered Memory"
- **Moumanti Podder**, Indian Institute of Science Education and Research, Pune, India
  "A model of market economics inspired by random walks with long memory"

IP38 — Mixing Behavior of Markovian Dynamics Room: 201

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Insuk Seo, Seoul National University, South Korea
Chair: Insuk Seo, Seoul National University, South Korea

- **Seonwoo Kim**, Yonsei University, South Korea
  "Transience time of the subcritical facilitated exclusion process"
- **Eric O. Endo**, New York University Shanghai, China
  "Phase Transition of the Long-Range Ising Models with Cell-Board External Fields"
- **Chiara Franceschini**, Università di Modena e Reggio Emilia, Italy
  "Degree-preserving conservative processes and a unified approach for their hydrodynamics"
- **Jungkyoung Lee**, Inha University, South Korea

"Mixing of the Curie–Weiss–Potts model"

IP36 — Recent Advances in Nonparametric Bayesian Learning

Room: 202

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Seung Jun Shin, Korea University, South Korea
Chair: Seung Jun Shin, Korea University, South Korea

- **Minwoo Chae**, Pohang University of Science and Technology, South Korea
  "Nonparametric Estimation of Undirected Graph Structures Using Diffusion Models"
- **Weining Shen**, University of California, Irvine, United States
  "Deep kernel learning based Gaussian processes for Bayesian image regression"
- **Seonghyun Jeong**, Yonsei University, South Korea
  "Bayesian Spatially Adaptive Triangulation"
- **Gyuhyeong Goh**, Kyungpook National U, South Korea
  "Bayesian infinite interactive fixed effects modeling for causal inference"

IP07 — Advances in Change Point Detection and Dependence Structures in High Dimensions

Room: 214

June 14 (Sunday) 13:00–14:40 Abstracts

Organizer: Qiwei Yao, The London School of Economics and Political Science, United Kingdom
Chair: Shakeel Gavioli-Akilagun, City University of Hong Kong, Hong Kong

- **Baojun Dou**, City University of Hong Kong, Hong Kong
  "Pairs Trading: Cointegration or Lack of It"
- **Zetai Cen**, University of Bristol, United Kingdom
  "Change Point Detection and Identification in Tensor Factor Models"
- **Yue Du**, Southwestern University of Finance and Economics, China
  "IDENTIFICATION AND ESTIMATION FOR MATRIX TIME SERIES CP-FACTOR MODELS"
- **Shakeel Gavioli-Akilagun**, City University of Hong Kong, Hong Kong
  "Optimal Online Change Detection via Random Fourier Features"

CS07 — Nonparametric and Semiparametric Statistical Methods

Room: 215

June 14 (Sunday) 13:00–14:40 Abstracts

Chair: Mr. Dennis Leung, University of Melbourne

- **Dennis Leung**, University of Melbourne
  "Berry-Esseen Theorems for the Asymptotic Normality of Incomplete U-Statistics with Bernoulli Sampling"
- **Taishi Kuzumoto**, University of Tokyo
  "Inference on Locally Adaptive Nonparametric Regression"
- **Deborshi Das**, Indian Statistical Institute, Delhi
  "Elephant Random Walks with Graph Based Shared Memory"
- **Daisuke Matsuno**, Tohoku University Graduate School of Information Sciences
  "Semiparametric Regression with Stagewise Minimization of the Empirical Risk"
- **Eun-Ji Lee**, Chungbuk National University
  "Simultaneous Estimation of Nonparametric Quantile Regression with B-Splines and Norm-Based Penalties"

14:50–16:30

DL06 — To be confirmed

Room: LT1A

June 14 (Sunday) 14:50–16:30 Abstracts

Chair: Xiaodan Fan, CUHK, Hong Kong

Distinguished Lecturer: Jun Liu, Harvard University, United States
"Conditional Generation via Diffusion, Flow, and Schrodinger Bridges"

- **Ke Deng**, Tsinghua University, China
  "Adaptive Kernel Density Estimation with Pre-training"
- **Minsuk Shin**, Yonsei University, South Korea
  "Amortizing Bootstrapped Nonparametric Maximum Likelihood Estimator"

DL04 — Minimax Optimality in Online Learning

Room: LT1B

June 14 (Sunday) 14:50–16:30 Abstracts

Chair: Enno Mammen, Heidelberg University, Germany

Distinguished Lecturer: Alexandre B. Tsybakov, ENSAE, France
"Continuum bandit, gradient-free stochastic optimization and nonparametric regression"

- **Arya Akhavan**, University of Oxford, United Kingdom
  "Non-stationary bandit convex optimization: A comprehensive study"
- **Alexandra Carpentier**, University of Potsdam, Germany
  "A simple and improved algorithm for noisy, convex, zeroth-order optimisation"

IP01 — Recent Development on High-Dimensional Data Modeling

Room: 209A

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Runze Li, The Pennsylvania State University, United States
Chair: Xin Tong, The University of Hong Kong, Hong Kong

- **Yingying Li**, The Hong Kong University of Science and Technology, Hong Kong
  "Site Percolation Network Models for Event-Driven Systems"
- **Jingyuan Liu**, Xiamen University, China
  "LLM-Powered Deep Panel Modeling with Application to Regional CPI Prediction"
- **Zhanrui Cai**, The University of Hong Kong, Hong Kong
  "A Statistical Framework for Alignment with Biased AI Feedback"
- **Yei Eun Shin**, Seoul National University, South Korea
  "Dynamic Network Modeling for the Spatiotemporal Progression of High-Dimensional Data"

IP19 — Recent Advances in High-dimensional Statistics

Room: 209B

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Mohamed Ndaoud, ESSEC Business School, France
Chair: Shu-Chin Lin, National Taiwan University, Taiwan

- **Dong Xia**, The Hong Kong University of Science and Technology, Hong Kong
  "TBD"
- **Peter Radchenko**, The University of Sydney, Australia
  "Extracting Interpretable Models from Tree Ensembles"
- **Anderson Ye Zhang**, University of Pennsylvania, United States
  "Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery"
- **Guillaume Braun**, RIKEN, Japan
  "Learning Dynamics of Phase Retrieval under Power-Law Data"

IP22 — Statistical Modeling with Deep Learning and Biomedical Applications

Room: 203

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Yuling Jiao, Wuhan University, China / Daewoo Pak, Yonsei University, South Korea
Chair: Yuling Jiao, Wuhan University, China

- **Lican Kang**, Wuhan University, China
  "TBD"

- **Ziyuan Chen**, Peking University, China
  "Deep Semiparametric Partial Differential Equation Models"
- **Daewoo Pak**, Yonsei University, South Korea
  "Genome-Wide Identification of Survival-Associated Genetic Variants under Interval Censoring via the Cox Model"
- **Chi Hyun Lee**, Yonsei University, South Korea
  "TBD"

IP41 — Stein's Method and Asymptotic Theory

Room: 201

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Xiao Fang, The Chinese University of Hong Kong, Hong Kong
Chair: Xiao Fang, The Chinese University of Hong Kong, Hong Kong

- **Adrian Röllin**, National University of Singapore, Singapore
  "Interplay of vertex and edge dynamics for dense random graphs"
- **Lihu Xu**, Michigan State University, United States
  "Quantitative bound for entropic CLT"
- **Wenkai Xu**, University of Warwick, United Kingdom
  "Stein discrepancies for testing and model assessments"
- **Yuta Koike**, The University of Tokyo, Japan
  "High-dimensional third order Edgeworth expansion by Stein's method"

IP35 — Statistical Methods for High-dimensional and Dependent Data

Room: 202

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Won Chang, Seoul National University, South Korea
Chair: Won Chang, Seoul National University, South Korea

- **Jun Song**, Korea University, South Korea
  "Sufficient Dimension Reduction via Dependence–Independence Classification"
- **Kyongwon Kim**, Yonsei University, South Korea
  "Learning causal graphs via nonlinear sufficient dimension reduction"
- **Junho Yang**, Academia Sinica, Taiwan
  "Denoising spatial point pattern data"
- **Byungwon Kim**, Kyungpook National University, South Korea
  "A Missing Value Imputation Method for High-Dimensional Tabular Data Using DeepInsight and Image Inpainting"

IP02 — Nonparametric Analysis of Euclidean and Non-Euclidean Data

Room: 214

June 14 (Sunday) 14:50–16:30 Abstracts

Organizer: Davy Paindaveine, Université libre de Bruxelles, Belgium
Chair: Sungkyu Jung, Seoul National University, South Korea

- **Andrea Meilan Vila**, Universidad Carlos III de Madrid, Spain
  "Quasi-likelihood estimation for semiparametric circular regression models"
- **Gaspard Bernard**, Academia Sinica, Taiwan
  "Testing for sphericity using spatial signs under elliptical directions"
- **Giacomo Francisci**, University of Trento, Italy
  "Data depth and dimension reduction"
- **Stanislav Nagy**, Charles University, Czech Republic
  "Computing depth for directional data"

CS04 — Classification, Variable Selection, and Deep Learning

Room: 215

June 14 (Sunday) 14:50–16:30 Abstracts

Chair: Dr. Bradley Rava, University of Sydney Business School

- **Bradley Rava**, University of Sydney Business School
  "Ask for More Than Bayes Optimal: A Theory of Indecisions for Classification"
- **Dongha Kim**, Sungshin Women's University
  "Anomaly Detection by Exploting Training Dynamics in Deep Generative Models"
- **Hirofumi Ota**, University of Tokyo
  "Fixed-Level Calibration of the Cauchy Combination Test"
- **Tao He**, San Francisco State University
  "Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data"

16:50–18:30

DL21 — Advancing Spatial Statistical Inference Through Machine Learning, AI, and Modern Computational Methods

Room: LT1A

June 14 (Sunday) 16:50–18:30 Abstracts

Chair: Soutir Bandyopadhyay, Colorado School of Mines, United States

Distinguished Lecturer: Douglas Nychka, Colorado School of Mines, United States
"Mining AI for statistical computation."

- **Bo Li**, Washington University in St. Louis, United States
  "Spatially Varying Deep Functional Neural Network: Application in Large-Scale Crop Yield Prediction"
- **Yeseul Jeon**, Texas A&M University, United States
  "Uncertainty-Aware Neural Multivariate Geostatistics"

IP62 — Frontiers in Statistical Modeling for Complex Data Structures

Room: LT1B

June 14 (Sunday) 16:50–18:30 Abstracts

Organizer: Ming-Yen Cheng, Hong Kong Baptist University, Hong Kong
Chair: Le Zhou, Hong Kong Baptist University, Hong Kong

- **Yuhong Yang**, Tsinghua University, China
  "On Mixture of Experts and Local Model Averaging"
- **Mengying You**, Shanghai University of International Business and Economics, China
  "Low-Rank Spatio-Temporal State Space Models with Structured Spatial Covariance"
- **Long Feng**, Nankai University, China
  "High dimensional alpha test for linear factor pricing model with L_q-norm"
- **Xu Guo**, Beijing Normal University, China
  "Inference of high-dimensional weak instrumental variable regression models without ridge-regularization"

IP48 — Markov Chain Simulation

Room: 209A

June 14 (Sunday) 16:50–18:30 Abstracts

Organizer: Ajay Jasra, The Chinese University of Hong Kong-Shenzhen, China
Chair: Ajay Jasra, The Chinese University of Hong Kong-Shenzhen, China

- **Michael Choi**, National University of Singapore, Singapore
  "Group-averaged Markov chains II: tuning of group action in finite state space"
- **Cosme Louart**, The Chinese University of Hong Kong-Shenzhen, China
  "Conditions for a Central Limit Theorem for Regularized M-Estimators with General Convex Losses"
- **Sheng Jiang**, The Chinese University of Hong Kong-Shenzhen, China
  "Errors-in-variables Gaussian Processes for Mixed-input Regression"

IP50 — Asymptotic Statistics — Room: 209B

June 14 (Sunday) 16:50–18:30 Abstracts

Organizer: Yunyi Zhang, The Chinese University of Hong Kong-Shenzhen, China
Chair: Yunyi Zhang, The Chinese University of Hong Kong-Shenzhen, China

- **Yunyi Zhang**, The Chinese University of Hong Kong-Shenzhen, China
  "Quadratic forms of high-dimensional non-stationary time series: Theory and Applications"
- **Nan Zou**, Macquarie University, Australia
  "Bootstrap for Dynamical Systems"
- **Yuqian Zhang**, Renmin University of China, China
  "Data integration using covariate summaries from external sources"
- **Yaoming Zhen**, The Chinese University of Hong Kong-Shenzhen, China
  "Probabilistic PCA on tensors"

PS — PS —Poster Session — Room: 203

June 14 (Sunday) 16:50–18:30

IP46 — High-dimensional Inference and Dependent Data Analysis — Room: 201

June 14 (Sunday) 16:50–18:30 Abstracts

Organizer: Shinpei Imori, Hiroshima University, Japan
Chair: Ching-Kang Ing, National Tsing Hua University, Taiwan

- **Shinpei Imori**, Hiroshima University, Japan
  "Greedy algorithms in high-dimensional linear regression models with group structure"
- **Yan Liu**, Waseda University, Japan
  "Testing for covariance structures in high-dimensional time series"
- **Hsueh-Han Huang**, Academia Sinica, Taiwan
  "High-dimensional transfer learning using greedy algorithms"
- **Valentin Patilea**, CREST & ENSAI, France
  "Testing the mean of multivariate random functions"

IP13 — Innovations in Bayesian Nonparametrics — Room: 202

June 14 (Sunday) 16:50–18:30 Abstracts

Organizer: Filippo Ascolani, Duke University, United States
Chair: Edwin Fong, The University of Hong Kong, Hong Kong

- **Li Ma**, The University of Chicago, United States
  "Assessing generative models through density ratio learning"
- **Igor Prünster**, Bocconi University, Italy
  "Multivariate species sampling models"
- **Junyi Zhang**, The Education University of Hong Kong, Hong Kong
  "Hierarchical Modelling in Bayesian Factor Analysis"

CS10 — Semiparametric Regression and Censored Data Analysis — Room: 214

June 14 (Sunday) 16:50–18:30 Abstracts

Chair: Dr. Sangbum Choi, Korea University

- **Zhi Yang Tho**, The Australian National University
  "Joint Mean and Correlation Regression Models for Multivariate Data"
- **Sangbum Choi**, Korea University
  "Identifiability and Inference of Semiparametric Copula-Based Quantile Regression under Dependent Censoring"
- **Arun Kaushik**, Banaras Hindu University, Varanasi, India

"Estimation of the Generalized Process Capability Index C_{pyk} under Generalized Progressive Hybrid Censoring using Maximum Likelihood, Robust and Bayesian Approaches"

- **Xiaoxi Zhang**, Seattle University
  "A Flexible Test for Comparing Two Hazard Rate Functions with Arbitrary Differences"
- **Charles Zhao**, University of North Carolina - Chapel Hill
  "Generalized Fiducial Method for Topic Modeling"

CS09 — Bayesian Inference, Monte Carlo, and Statistical Estimation

Room: 215

June 14 (Sunday) 16:50–18:30 **Abstracts**

Chair: Mr. Ka Lok Lam, University of California, Santa Barbara

- **Pushkar Mohan Kale**, National University of Singapore, Singapore
  "Moment Constrained Cutting Feedback for Modular Bayesian Models"
- **Ka Lok Lam**, University of California, Santa Barbara
  "Probabilistic representation and Monte Carlo method for nonlinear Dirac equations via telegraph particles"
- **Linus David Fromm**, University of Otago
  "Efficiency of MCMC Samplers for Discrete Inverse Problems"
- **Jyun-Yu Chen**, Academia Sinica
  "Efficient and Interpretable Mixtures of Experts: Statistical Inference, Initialization, and Applications"

19:00 **Conference Banquet**

June 15 (Monday)

09:00–10:40

DL07 — To be confirmed
Room: LT1A

June 15 (Monday) 09:00–10:40 Abstracts

Distinguished Lecturer: George Michailidis, University of California, Los Angeles, United States
"TBD"

- **Yao Zheng**, University of Connecticut, United States
  "TBD"

DL08 — To be confirmed
Room: LT1B

June 15 (Monday) 09:00–10:40 Abstracts

Distinguished Lecturer: Xiao Wang, Purdue University, United States
"Coreset-Induced Flow Matching"

- **Yongdai Kim**, Seoul National University, South Korea
  "Uncertainty-adaptive Feedback Guidance for Improved Image Generation with Diffusion Models"
- **Yixuan Qiu**, Shanghai University of Finance and Economics, China
  "GPU-Accelerated Solver for Entropic-Regularized Optimal Transport"

IP56 — Modern Nonparametric Approaches for Dependent Data Analysis
Room: 209A

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Soutir Bandyopadhyay, Colorado School of Mines, United States
Chair: Bo Li, Washington University in St. Louis, United States

- **Chae Young Lim**, Seoul National University, South Korea
  "Exploring Spatial Dynamics in Regression Coefficients: A Bayesian Regularization Method with Clustering"
- **Arindam Chatterjee**, Indian Statistical Institute, Delhi, India
  "Predicting network summary statistics through network sampling: some rigorous results under induced and egocentric network formation"
- **Soudeep Deb**, Indian Institute of Management Bangalore, India
  "Nonparametric regression of spatio-temporal data using infinite-dimensional covariates"
- **Soutir Bandyopadhyay**, Colorado School of Mines, United States
  "Frequency Domain Resampling for Gridded Spatial Data"

IP15 — Statistical Machine Learning for High-dimensional Neuroimaging Time Series
Room: 209B

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Ali Shojaie, University of Washington, United States
Chair: Gongjun Xu, University of Michigan, United States

- **Hernando Ombao**, King Abdullah University of Science and Technology, Saudi Arabia
- **Michele Guindani**, University of California, Los Angeles, United States
  "Decoding Neuronal Ensembles from Spatially-Referenced Calcium Traces: A Bayesian Semiparametric Approach"
- **Eardi Lila**, University of Washington, United States
  "Biophysics-informed deep operator learning for electrophysiological source reconstruction"

IP58 — Frontiers in Bayesian Learning and Inference

Room: 203

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Susan Wei, Monash University, Australia
Chair: Minwoo Chae, Pohang University of Science and Technology, South Korea

- **Yingzhen Li**, Imperial College London, United Kingdom
  "LLMs as implicit/predictive Bayesian models: algorithmic frontiers"
- **David Frazier**, Monash University, Australia
  "Predictive Bayesian Inference on Population Functionals"
- **Susan Wei**, Monash University, Australia
  "Pretrained Transformers as In-Context Bayesian Learners: Implications for Uncertainty Quantification"

IP34 — Recent Advances in Stochastic Processes

Room: 201

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Teppei Ogihara, The University of Tokyo, Japan
Chair: Teppei Ogihara, The University of Tokyo, Japan

- **Junichiro Yoshida**, The University of Tokyo, Japan
  "Estimation Error and Hypothesis Testing for Non-identifiable Models, with Applications to Machine Learning"
- **Jie Yen Fan**, Monash University, Australia
  "Estimation in age-and-population-dependent models"
- **Tetsuya Takabatake**, The University of Osaka, Japan
  "Optimal Estimation for General Gaussian Processes in the Frequency Domain"
- **Yasutaka Shimizu**, Waseda University, Japan
  "Joint estimation for mean functions and covariance kernels in Gaussian processes"

IP54 — Sampling and Optimization in Modern Data Science Problems

Room: 202

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Debdeep Pati, University of Wisconsin–Madison, United States
Chair: Lancelot Fitzgerald James, The Hong Kong University of Science and Technology, Hong Kong

- **Ning (Patricia) Ning**, Texas A&M University, United States
  "Robust Iterative Learning Hidden Quantum Markov Models"
- **Yuchen Wu**, Cornell University, United States
  "On the Robustness of Distribution Support under Diffusion Guidance"
- **Rong Tang**, The Hong Kong University of Science and Technology, Hong Kong
  "Robust Bayesian Inference on Riemannian Submanifold"
- **Runmin Wang**, Texas A&M University, United States
  "Change-point detection in high-dimensional time series using MOSUM"

IP55 — Modern Perspectives on Causal Inference

Room: 214

June 15 (Monday) 09:00–10:40 Abstracts

Organizer: Sivaraman Balakrishnan, Carnegie Mellon University, United States
Chair: Yifan Cui, Zhejiang University, China

- **Raaz Dwivedi**, Cornell University, United States
  "TBD"
- **Shuangning Li**, The University of Chicago, United States
  "Covariate Adjustment Cannot Hurt: Treatment Effect Estimation under Interference with Low-Order Outcome Interactions"
- **Rajarshi Mukherjee**, Harvard University, United States
  "Inference in high-dimensional linear mediation models under proportional asymptotics"
- **Debraj Das**, Indian Institute of Technology Bombay, India
  "Asymptotic Theory of K-fold Cross-validation in Lasso and the Validity of Bootstrap"

CS03 — High-Dimensional Inference, Factor Models, and Latent Variables — Room: 215

June 15 (Monday) 09:00–10:40 Abstracts

Chair: Ms. Yuqi Zhang, University of Bristol

- **Yuqi Zhang**, University of Bristol
  "Multiscale Detection of Multiple Change Points in High-Dimensional Factor Models"
- **Kyoowon Kim**, Seoul National University
  "Testing and Segmentation of Joint and Individual Components in Integrative Multi-Source Factor Models"
- **Zhining Wang**, The Australian National University
  "Simultaneous Inference for Latent Variable Predictions in Factor Analytic Models"
- **Giheon Seong**, Seoul National University
  "James–Stein Estimation of Spiked Eigenvectors Under the Generalized Spiked Population Model"
- **Kanta Naito**, Graduate School of Information Sciences
  "Aspects of High-Dimensional Kernel Density Estimation: Bandwidth-Induced Bifurcations and Their Estimation"

11:00–12:40

DL17 — Change Point Detection and Its Modern Applications — Room: LT1A

June 15 (Monday) 11:00–12:40 Abstracts

Chair: Yingcun Xia, National University of Singapore, Singapore

Distinguished Lecturer: Jialiang Li, National University of Singapore, Singapore
"Change-point Detection and Its Modern Applications"

- **Alex Luedtke**, University of Washington, United States
  "Simplifying debiased inference via automatic differentiation and probabilistic programming"
- **Jessica Li**, University of California, Los Angeles, United States
  "TwinPoSI: A Synthetic Data-Based Method for Valid and Powerful Post-Selection Inference"

DL20 — Mathematical Underpinnings of Distributed Inference: From Theory to Real-world Applications — Room: LT1B

June 15 (Monday) 11:00–12:40 Abstracts

Chair: Ying Zhang, University of Nebraska Medical Center, United States

Distinguished Lecturer: Yong Chen, University of Pennsylvania, United States
"Mathematical Underpinnings of Distributed Inference: From Theory to Real-world Applications"

- **Jingmei Qiu**, University of Delaware, United States
  "TBD"
- **Yudong Wang**, National University of Singapore, Singapore
  "TT-MOSAiC: A Scalable Tensor-Train-Powered Framework for One-Shot and Lossless Federated Learning"

IP63 — Young Researchers' Session — Room: 209A

June 15 (Monday) 11:00–12:40 Abstracts

Organizer: Keisuke Yano, The Institute of Statistical Mathematics, Japan
Chair: Keisuke Yano, The Institute of Statistical Mathematics, Japan

- **Minwoo Kim**, Seoul National University, South Korea
  "Enhancing Differentially Private Mechanisms via Empirical Bayes"
- **Tetsuya Umino**, University of Tsukuba, Japan
  "Automatic Sparse Estimation of High-Dimensional Mean Vectors in Multisample Settings"
- **Haruka Yoshida**, Yokohama National University, Japan
  "Multiple Effect Restoration for Measurement and Confounding Bias in Causal Inference"
- **Yuki Takazawa**, The University of Tokyo, Japan

"Robust Species Tree Inference Using Modes of Quartet Topologies in Tree Space"

IP67 — Advances in Flexible and Adaptive Statistical Inference

Room: 209B

June 15 (Monday) 11:00–12:40 Abstracts

Organizer: Hoseung Song, Korea Advanced Institute of Science and Technology, South Korea
Chair: Hoseung Song, Korea Advanced Institute of Science and Technology, South Korea

- **Doudou Zhou**, National University of Singapore, Singapore
  "MATES: Multi-view Aggregated Two-Sample Test"
- **Kwangho Kim**, Korea University, South Korea
  "Toward Flexible and Efficient Counterfactual Density Estimation"
- **Jingru Zhang**, Fudan University, China
  "Harmonizing Time-Varying Physical Activity Data Across Wearable Devices"
- **Jie Wang**, The Chinese University of Hong Kong-Shenzhen, China
  "Variable Selection for Kernel Two-Sample Tests"

IP72 — Bayesian Learning of Complex Structures

Room: 203

June 15 (Monday) 11:00–12:40 Abstracts

Organizer: Subhashis Ghoshal, North Carolina State University, United States
Chair: Igor Pruenster, Bocconi University, Italy

- **Pierre Alquier**, ESSEC Business School, Singapore
  "Rates of convergence in Bayesian meta-learning"
- **Jaeyong Lee**, Seoul National University, Korea
  "Eigenstructure inference for the high-dimensional covariance matrices with generalized shrinkage inverse-Wishart prior"
- **William Weimin Yoo**, Heriot-Watt University Malaysia, Malaysia
  "Learning Weights and Depth in Bayesian Neural Networks via Markov Chain Approximations"
- **Subhashis Ghoshal**, North Carolina State University, United States
  "Bayesian learning of relational graph in semiparametric high-dimensional time series"

IP39 — Innovative Methods in Survival Analysis and Survey Data Integration

Room: 201

June 15 (Monday) 11:00–12:40 Abstracts

Organizer: Xinyuan Song, The Chinese University of Hong Kong, Hong Kong
Chair: Xinyuan Song, The Chinese University of Hong Kong, Hong Kong

- **Jianguo Sun**, Southern University of Science and Technology, China
  "A new transfer learning estimation approach for failure time data"
- **Dipankar Bandyopadhyay**, Viginia Commonwealth University, United States
  "Rank estimation for the accelerated failure time model under partially interval-censored data"
- **Changbao Wu**, University of Waterloo, Canada
  "Data Integration with Non-Probability Survey Samples"
- **Yi Li**, University of Michigan, United States
  "Inference for the Relative Risk Functional in Deep Nonparametric Cox Models"

IP66 — Statistical Methods for Complex and Non-Euclidean Data

Room: 202

June 15 (Monday) 11:00–12:40 Abstracts

Organizer: Joonpyo Kim, Chung-Ang University, South Korea
Chair: Joonpyo Kim, Chung-Ang University, South Korea

- **Ho Yun**, École Polytechnique Fédérale de Lausanne, Switzerland
  "Spectral Gaps and Spherical Harmonics: A Directional Statistics Approach to DNA Flexibility"
- **Almond Stöcker**, École Polytechnique Fédérale de Lausanne, Switzerland
  "Kernel ridge regression for spherical responses"

- **Seungwoo Kang**, Sungkyunkwan University, South Korea
  "L_1 Prominence Measures for Directed Graphs"
- **Seoncheol Park**, Hanyang University, South Korea
  "Adaptive Boosting on Linear Networks"

CS01 — Bayesian and Likelihood-Based Inference

Room: 214

June 15 (Monday) 11:00–12:40 Abstracts

Chair: Dr. Weichang Yu, University of Melbourne

- **Weichang Yu**, University of Melbourne
  "Cutting Feedback in Misspecified Copula Models"
- **Yichen Zhu**, The University of Hong Kong
  "Vecchia Gaussian Processes: On Probabilistic and Statistical Properties"
- **Khue-Dung Dang**, University of Western Australia
  "Variational Approximate Penalized Credible Regions for Bayesian Grouped Regression"
- **Nelson Jinn-Yih Chua**, Australian National University
  "Asymptotic Results for Model Parameters and Random Effects Inference Using Gaussian Variational Approximations in Generalized Linear Mixed Models"
- **Shogo Kusano**, Kumamoto University
  "Quasi-Bayesian Information Criterion of SEM for Diffusion Processes"

CS13 — Applied Probability, Financial Statistics, and Cure Models

Room: 215

June 15 (Monday) 11:00–12:40 Abstracts

Chair: Mr. Hugh Entwistle, Macquarie University

- **Hugh Entwistle**, Macquarie University
  "On Optimal Stopping Problems with Random Supply"
- **Yuanhang Luo**, The Hong Kong Polytechnic University
  "Adaptive Debiased Lasso in High-dimensional Generalized Linear Models with Streaming Data"
- **Lovely Aisha Jamil**, American University of Sharjah
  "Understanding the Evolution of Kyle's Lambda on Digital Blockchain Assets"
- **Junyan Ye**, The Chinese University of Hong Kong
  "Martingale Duality Meets Statistical Learning: Deep Primal-Dual Bounds and Policy Learning for High-Dimensional Optimal Switching Problem"

Afternoon **Half-day Break**

June 16 (Tuesday)

09:00–10:40

DL13 — Discrete Random Models and Learning

Room: LT1A

June 16 (Tuesday) 09:00–10:40 Abstracts

Chair: Parthanil Roy, Indian Institute of Technology Bombay, India

Distinguished Lecturer: Antar Bandyopadhyay, Indian Statistical Institute, Delhi, India
"Interacting Urn Schemes"

- **Subhro Ghosh**, National University of Singapore, Singapore
  "Strongly correlated particle systems: a toolbox for machine intelligence"
- **Nathan Ross**, University of Melbourne, Australia
  "Detecting correlation in uniform attachment trees"

DL14 — Development of Causal Inference

Room: LT1B

June 16 (Tuesday) 09:00–10:40 Abstracts

Chair: Tomonari Sei, The University of Tokyo, Japan

Distinguished Lecturer: Manabu Kuroki, Yokohama National University, Japan
"The Evaluation of the Probablities of Potential Outcome Types from Statistical Data"

- **Yuta Kawakami**, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
  "Moments of Causal Effects"
- **Shunichiro Orihara**, Tokyo Medical University, Japan
  "Average Treatment Effect Estimation under Poor Overlap via Weighted Estimands"

IP29 — Recent Advances in Reinforcement Learning

Room: 209A

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Moulinath Banerjee, University of Michigan, United States
Chair: Moulinath Banerjee, University of Michigan, United States

- **Daniele Bracale**, University of Michigan, United States
  "Online Price Competition under Generalized Linear Demands"
- **Kihyuk Hong**, Korea Advanced Institute of Science and Technology, South Korea
  "Recent Advances in Offline Constrained Reinforcement Learning"
- **Shivaram Kalyanakrishnan**, Indian Institute of Technology Bombay, India
  "On-line Learning in Tree MDPs by Treating Policies as Bandit Arms"
- **Wen Sun**, Cornell University, United States
  "TBD"

IP12 — Experiment Design in the Modern Era

Room: 209B

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Kean Ming Tan, University of Michigan, United States
Chair: Jeong Min Jeon, Seoul National University, South Korea

- **Lixin Zhang**, Zhejiang GongShang University, China
  "Asymptotic Properties of Covariate Adaptive Randomization Procedures for Balancing Observed and Unobserved Covariates"
- **Wei Ma**, Renmin University of China, China
  "Covariate-adaptive design: An overview and recent advances"
- **Waverly Wei**, University of Southern California, United States

"Trace-Aware Routing for Cost-Effective Human–AI Collaborative Labeling"

- **Jingshen Wang**, University of California, Berkeley, United States
  "TBD"

IP21 — Innovations in Changepoint Detection

Room: 203

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Changliang Zou, Nankai University, China
Chair: Changliang Zou, Nankai University, China

- **Ying Yang**, Fudan University, China
  "Spatially Randomized Designs Can Enhance Policy Evaluation"
- **Wei Zhang**, Fudan University, China
  "Structural Change Detection in Dynamic Systems"
- **Chengde Qian**, Shanghai Jiao Tong University, China
  "Changepoint Detection in Complex Models: Cross-Fitting Is Needed"
- **Guanghui Wang**, Nankai University, China
  "ART: Distribution-Free and Model-Agnostic Changepoint Detection with Finite-Sample Guarantees"

IP52 — Recent Advancements in Semiparametric Methods for Complex Censored Outcomes

Room: 201

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Yu Gu, The University of Hong Kong, Hong Kong / Yangjianchen Xu, University of Waterloo, Canada
Chair: Yangjianchen Xu, University of Waterloo, Canada

- **Yu Gu**, The University of Hong Kong, Hong Kong
  "Semiparametric Functional Multi-State Models with Application to Alzheimer's Disease"
- **Yangjianchen Xu**, University of Waterloo, Canada
  "Robust inference for the Cox proportional hazards model with interval-censored data"
- **Kin Yau (Alex) Wong**, The Hong Kong Polytechnic University, Hong Kong
  "A unified two-step estimation approach for semiparametric models under two-phase sampling"

IP49 — Stochastic Processes and Statistics

Room: 202

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Kengo Kamatani, The Institute of Statistical Mathematics, Japan
Chair: Kengo Kamatani, The Institute of Statistical Mathematics, Japan

- **Masahiro Kurisaki**, RIKEN, Japan
  "Asymptotic Expansion of Nonlinear Filtering"
- **Hirofumi Shiba**, The Institute of Statistical Mathematics, Japan
  "Diffusive Scaling Limits & Early Diagnostics for Piecewise Deterministic Monte Carlo Samplers"
- **Teppei Ogihara**, The University of Tokyo, Japan
  "Asymptotically uniformly most powerful tests for diffusion processes with nonsynchronous observations"
- **Shogo Nakakita**, The University of Osaka, Japan
  "Dimension-free uniform concentration bound for logistic regression"

IP32 — Recent Advances in Theories and Methodologies for High-dimensional PCA

Room: 214

June 16 (Tuesday) 09:00–10:40 Abstracts

Organizer: Kazuyoshi Yata, University of Tsukuba, Japan
Chair: Kazuyoshi Yata, University of Tsukuba, Japan

- **Koji Tsukuda**, Kyushu University, Japan

"Multivariate allometric regression: methods and theory"

- **Kazuyoshi Yata**, University of Tsukuba, Japan
  "Asymptotic Properties of Automatic Sparse PCA for High-Dimensional Data and Its Applications"
- **Shao-Hsuan Wang**, National Central University, Taiwan
  "Bayesian sparse principal coordinates analysis with microbiome discoveries"

Discussant: Kento Egashira, Tokyo University of Science, Japan

CS02 — Causal Inference and Treatment Effects

Room: 215

June 16 (Tuesday) 09:00–10:40 Abstracts

Chair: Prof. Yuming Sun, William & Mary

- **Taehyeon Koo**, Columbia University
  "Distributionally Robust Synthetic Control: Ensuring Robustness Against Highly Correlated Controls and Weight Shifts"
- **Yuming Sun**, William & Mary
  "Estimating Heterogeneous Treatment Effects with Survival Outcomes via Deep Survival Learner"
- **Richard Guo**, University of Michigan
  "Hunt-and-test strategies for ML-powered hypothesis testing"
- **Ha-Young Shin**, Soongsil University
  "Treatment Effects on Hadamard Spaces"
- **Hisayuki Hara**, Kyoto University
  "Spatial Statistical Models for Obsidian Source Composition"

11:00–12:40

DL15 — Frontiers in Non-Euclidean Data Analysis

Room: LT1A

June 16 (Tuesday) 11:00–12:40 Abstracts

Chair: Chae Young Lim, Seoul National University, South Korea

Distinguished Lecturer: Sungkyu Jung, Seoul National University, South Korea
"Generalized Frechet Means and their applications"

- **Zhenhua Lin**, National University of Singapore, Singapore
  "Fréchet Single-Index Regression: Regularization, Estimation, and Optimality"
- **Jongmin Lee**, Pusan National University, Korea
  "Huber means on Riemannian manifolds"

DL19 — On Sinkhorn Semigroups and Related Fields

Room: LT1B

June 16 (Tuesday) 11:00–12:40 Abstracts

Chair: Ajay Jasra, The Chinese University of Hong Kong-Shenzhen, China

Distinguished Lecturer: Pierre Del Moral, INRIA, France & The Chinese University of Hong Kong-Shenzhen, China
"New Trends in the Stability Analysis of Sinkhorn Semigroups"

- **Xin Tong**, National University of Singapore, Singapore
  "Wasserstein gradient flow for optimal distribution decomposition"
- **Ajay Jasra**, The Chinese University of Hong Kong-Shenzhen, China
  "New Trends in the Stability of Sinkhorn Semigroups"

IP53 — Machine Learning-informed Inference and Decision Making

Room: 209A

June 16 (Tuesday) 11:00–12:40 Abstracts

Organizer: Xiaowu Dai, University of California, Los Angeles, United States

Chair: Xiaowu Dai, University of California, Los Angeles, United States

- **Yichen Zhang**, Purdue University, United States
  "A Sparse Learning Framework for the High-Dimensional Newsvendor"
- **Sriram Sankararaman**, University of California, Los Angeles, United States
  "Recent advances and challenges in imputation methods for large-scale biomedical data"
- **Heng Lian**, City University of Hong Kong, Hong Kong
  "Distributed semi-supervised inference for generalized linear models with block-wise missing covariates"
- **Qiyang Han**, Rutgers University, United States
  "Bandit algorithms: Precise dynamics and statistical inference"

IP64 — Advances in Statistical and Machine Learning Methods for Biomedical Applications

Room: 209B

June 16 (Tuesday) 11:00–12:40 Abstracts

Organizer: Junsouk Choi, Korea University, South Korea
Chair: Jun Song, Korea University, South Korea

- **Sehwan Kim**, Ewha Womans University, South Korea
  "Self-Consistent Equation-guided Neural Networks for Censored Time-to-Event Data"
- **Taehwa Choi**, Sungshin Women's University, South Korea
  "Interval-censored linear quantile regression"
- **Guanxun Li**, Beijing Normal University at Zhuhai, China
  "E-value Aggregation via Data-Dependent Weighting and Its Application to Omics-Wide Differential Analysis"
- **Junsouk Choi**, Korea University, South Korea
  "Semi-supervised Spatial Topic Modeling for Discovery of Multicellular Spatial Tissue Structures in Multiplex Imaging"

IP68 — Statistical and Optimization Perspectives of Generative Models

Room: 203

June 16 (Tuesday) 11:00–12:40 Abstracts

Organizer: Yao Xie, Georgia Institute of Technology, United States
Chair: Yoonkyung Lee, The Ohio State University, United States

- **Masashi Sugiyama**, The University of Tokyo, Japan
  "Recent Advances in Reward Modeling for Reinforcement Learning"
- **Xiuyuan Cheng**, Duke University, United States
  "Minimax learning in Wasserstein space via neural transport maps"
- **Yuting Wei**, University of Pennsylvania, United States
  "Dimension-Free Convergence of Diffusion Models for Approximate Gaussian Mixtures"
- **Yuan Yao**, The Hong Kong University Science and Technology, Hong Kong
  "TBD"

IP71 — IMS New Researchers Invited Session

Room: 201

June 16 (Tuesday) 11:00–12:40 Abstracts

Organizer: Armeen Taeb, University of Washington, United States
Chair: Alexander Giessing, National University of Singapore, Singapore

- **Xiaozhu Zhang**, University of Washington, United States
  "Convex Mixed-Integer Programming for Causal Additive Models with Optimization and Statistical Guarantees"
- **Charlie Wolock**, University of Rochester, United States
  "Leveraging machine learning to estimate survival curves with current status data"
- **Julien Laurendeau**, École Polytechnique Fédérale de Lausanne, Switzerland
  "New guarantees for optimal regimes in the presence of unmeasured confounding"
- **Jin-Hong Du**, The University of Hong Kong, Hong Kong
  "Seeing Through Correlations: Disentangled Feature Importance"

CS05 — Time Series, Stochastic Processes, and Financial Statistics

Room: 202

June 16 (Tuesday) 11:00–12:40 Abstracts

Chair: Miss. Ziyuan Zhang, Purdue University

- **Gehui Zhang**, Southwest Petroleum University
  "Multivariate Stochastic Volatility with Informative Missingness"
- **Boyuan Ning**, Waseda University
  "Estimation of the Elasticity for CKLS Model from High-Frequency Observations"
- **Edward Hill**, Queen Mary University of London
  "Nonstationarity Extended Whittle Estimation of Cyclical Time Series"
- **Jiazhen Xu**, Macquarie University
  "Spherically Embedded Time Series with Unknown Trend and Periodic Components"
- **Ziyuan Zhang**, Purdue University
  "Data Driven Asset Pricing"

CS06 — Survival Analysis, Clinical Trials, and Biostatistics

Room: 214

June 16 (Tuesday) 11:00–12:40 Abstracts

Chair: Prof. Yanqing Yi, Memorial University of Newfoundland

- **HYEON SEOK Oh**, Korea University
  "Identifiability in Semiparametric-Parametric Frailty Models with Dependent Competing Risks"
- **Shanpeng Li**, City of Hope
  "Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Competing Risks Time-to-Event Data: Applications to Mega-Scale Health Research"
- **Yanqing Yi**, Memorial University of Newfoundland
  "Optimal Adaptive Randomized Clinical Trial with Covariates"
- **Abhimanyu Singh Yadav**, Banaras Hindu University
  "On Generalized Adaptive Progressive First-Failure Censoring Scheme with Optimal Design"
- **Yuji Komiyama**, Tohoku University
  "A Finite-Time-Horizon Mixture Cure Model with Application to Online Marketplace Data"

CS11 — Change Point Detection, High-Dimensional Time Series, and Functional Data

Room: 215

June 16 (Tuesday) 11:00–12:40 Abstracts

Chair: Ms. DEBANJANA DATTA, Indian Statistical Institute

- **Dylan Dijk**, University of Bristol
  "Tail-Robust Change Point Detection in High-Dimensional Linear Regression with Non-Sparse Structures"
- **Debanjana Datta**, Indian Statistical Institute
  "Detection of Structural Shifts in Functional Time Series"
- **Jaesung Park**, Seoul National University
  "Principal Component Analysis for Zero-Inflated Compositional Data"
- **Namgil Lee**, Kangwon National University
  "A Unified Framework for Shrinkage Estimation of High-Dimensional Vector Autoregressive Models with the R Package VARshrink"
- **Mingxu LI**, Beijing Normal Hong Kong Baptist University
  "A GARCH-MIDAS-CJ Model: Integrating Jump Decomposition into Multiplicative Component Volatility"

13:30–15:10

IP43 — Advances in Deep Learning and Kernel Learning Methods

Room: LT1A

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Cheng Li, National University of Singapore, Singapore
Chair: Cheng Li, National University of Singapore, Singapore

- **Won Chang**, Seoul National University, South Korea
  "Deep Normalizing Flow Methods for Fast Bayesian Inference"
- **Yuqi Gu**, Columbia University, United States
  "Discrete Causal Representation Learning"
- **Qian Lin**, Tsinghua University, China
  "The adaptive feature program"
- **Feng Ruan**, Northwestern University, United States
  "A Theory of Feature Learning in Kernel Models"

IP31 — Optimal Transport and Beyond

Room: LT1B

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Takeru Matsuda, The University of Tokyo & RIKEN, Japan
Chair: Takeru Matsuda, The University of Tokyo & RIKEN, Japan

- **Ting-Kam Leonard Wong**, University of Toronto, Canada
  "Shape constrained density estimation with Wasserstein projection"
- **Matthew Thorpe**, University of Warwick, United Kingdom
  "Laplace Learning in Wasserstein Space"
- **Tin Lok James Ng**, Trinity College Dublin, Ireland
  "Bayesian Spatially Varying Regression with Optimal Transport"
- **Yoshikazu Terada**, The University of Osaka, Japan
  "A New Perspective on Matrix Decomposition Factor Analysis"

IP57 — Recent Advances in Variational Inference

Room: 209A

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Minh-Ngoc Tran, The University of Sydney, Australia
Chair: Kyongwon Kim, Yonsei University, South Korea

- **Kamélia Daudel**, ESSEC Business School, France
  "Importance Weighted Variational Inference without the Reparameterization Trick"
- **Dario Draca**, The University of Sydney, Australia
  "Inversion-Free Natural Gradient Descent on Riemannian Manifolds"
- **Cheng Zhang**, Peking University, China
  "Provable Sample-Efficient Transfer Learning Conditional Diffusion Models via Representation Learning"
- **Minh-Ngoc Tran**, The University of Sydney, Australia
  "Bures-Wasserstein Importance-Weighted Evidence Lower Bound: Exposition and Applications"

IP20 — Statistical Estimation and Detection in Complex Modeling

Room: 209B

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Wei Zhong, Xiamen University, China
Chair: Wei Zhong, Xiamen University, China

- **Haojie Ren**, Shanghai Jiao Tong University, China
  "TBD"
- **Shunxing Yan**, Peking University, China
  "Semiparametric M-estimation with Overparameterized Neural Networks"
- **Xiaodong Yan**, Xian Jiaotong University, China
  "AI Safety: Statistical Detection for Silent Data Corruption during Large-scale Model Training and Reasoning"
- **Xuening Zhu**, Fudan University, China

"TBD"

IP51 — Advanced Machine Learning Methods for Challenges in Biomedical Data Room: 203

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Ran Dai, University of Nebraska Medical Center, United States
Chair: Cheng Zheng, University of Nebraska Medical Center, United States

- **Changhu Wang**, University of California, Los Angeles, United States
  "Nullstrap: A Simple, High-Power, and Fast Framework for FDR Control in Variable Selection for Diverse High-Dimensional Models"
- **Mladen Kolar**, University of Southern California, United States
  "SMART: A Spectral Transfer Approach to Multi-Task Learning"
- **Hongyuan Cao**, Florida State University, United States
  "Statistical methods for fine mapping in admixed populations"
- **Ran Dai**, University of Nebraska Medical Center, United States
  "Controlling FDR in selecting group-level simultaneous signals from multiple data sources with application to the National COVID Collaborative Cohort data"

IP42 — Advances in Random Graph Theory Room: 201

June 16 (Tuesday) 13:30–15:10 Abstracts

Organizer: Adrian Röllin, National University of Singapore, Singapore
Chair: Adrian Röllin, National University of Singapore, Singapore

- **Yuanfei Huang**, Asia-Pacific Center for Theoretical Physics, South Korea
  "The SIR epidemic on a dynamic Erd{\H{o}}s-Rényi random graph"
- **Xiao Fang**, The Chinese University of Hong Kong, Hong Kong
  "Conditional central limit theorems for exponential random graphs"
- **Rajat Hazra**, Leiden University, The Netherlands
  "Voter Model on random graphs"
- **Gianmarco Bet**, University of Florence, Italy
  "Localized geometry detection in scale-free random graphs"

CS08 — Conformal Prediction, Reinforcement Learning, and Modern Statistical Methods Room: 202

June 16 (Tuesday) 13:30–15:10 Abstracts

Chair: Dr. Tongyu Li, National University of Singapore

- **Chihoon Lee**, Seoul National University
  "Predicting Current Outcomes From Historical Survey Data With Weighted Conformal Prediction"
- **Wenbo Jing**, City University of Hong Kong
  "Knowledge Transfer in Batch Q Learning"*
- **Soohyun Ahn**, Ajou University
  "Conformal Monitoring of Complex Data with Finite-Sample Validity"
- **Yongjae Kim**, Seoul National University
  "An Association Measure for Mixed-Types Variables"
- **Tongyu Li**, National University of Singapore
  "Statistical Inference on Gradient Flows"

CS12 — Biostatistics, Genomics, and Multiple Testing Room: 214

June 16 (Tuesday) 13:30–15:10 Abstracts

Chair: Mr. Ninh Tran, University of Melbourne

- **Hanning Chen**, University of Melbourne
  "Moderated t-Covariates Improve Gene Ranking in Large-Scale Differential Expression Studies with Small

Sample Sizes"

- **Ninh Tran**, University of Melbourne
  "A Covariate-Adaptive Test for Replicability Across Multiple Studies with False Discovery Rate Control"
- **Armeen Taeb**, University of Washington
  "Consensus Tree Estimation with False Discovery Rate Control via Partially Ordered Sets"
- **Qin Shao**, University of Toledo
  "Innovative Imputation Strategies for Time Series Data from Wearable Devices and Mobile Applications"

CS14 — Latent Variable Models, Identifiability, and Complex Dependence Structures Room: 215

June 16 (Tuesday) 13:30–15:10 **Abstracts**

Chair: Dr. Jimin Kim, Seoul National University

- **Chengyu Cui**, University of Michigan
  "Beyond Vintage Rotation: Bias-Free Sparse Representation Learning with Oracle Inference"
- **Jiayi Huang**, University of Virginia
  "How Does Missing Data Affect Latent Transition Analysis? A Monte Carlo Study"
- **Chengzhu Huang**, Columbia University
  "A General Recipe for Generalized Latent Factor Models: Missingness, Implicit Regularization, and Inference"
- **Mengqi Lin**, University of Michigan
  "Characterizing Identifiability in Boolean Graphical Models"
- **Jimin Kim**, Seoul National University
  "Conditional Copula Networks for Synthetic Tabular Data with Complex Dependencies"

Talk Abstracts

DL02 — Causal Inference

June 13 (Saturday) 10:50–12:30 **Room:** LT1A Program

Jane-Ling Wang, University of California, Davis, United States

Quantile Treatment Effect Estimation from Censored Lifetime Data

Many economic and medical studies aim to estimate the effect of treatments on lifetimes. Estimating the average treatment effect is challenging because lifetime data are typically right censored, which prevents accurately estimating the upper tail of the survival function. Existing approaches thus bypass the average treatment effect and instead focus on estimating the hazard ratio or restricted mean lifetime. Such approaches often rely on strong model assumptions that may be violated in real studies. In this paper, we advocate the use of quantile treatment effects for censored time-to-event outcomes. This approach not only provides more comprehensive insights across multiple quantiles but also avoids the challenges of average treatment effect estimation and other drawbacks of existing approaches. In addition, it offers robustness against outliers and in heavy-tailed settings. Based on the Neyman-Rubin potential outcomes framework and martingale processes, we propose a model-free approach to estimate quantile treatment effects by solving inverse probability weighted estimating equations. The proposed estimator is proven to be uniformly consistent and weakly convergent at the parametric root-n rate. Simulations confirm its finite-sample performance in complex settings, and we demonstrate its application using real data from an employment experiment and an AIDS clinical trial.

Qixian Zhong, Xiamen University, China

Generative Doubly Robust Estimation for General Treatment Effects

This paper introduces a unified framework for doubly robust (DR) estimation of a broad class of causal functionals, including average, quantile, and asymmetric least squared treatment effects, as well as their conditional counterparts. While DR estimators are well-established for average treatment effects, their development for distributional parameters like quantile treatment effects has not yet been investigated. We bridge this gap by integrating conditional generative models into a loss-based estimating framework. Our approach uses generative models to synthesize counterfactual samples, defines a target loss whose minimizer corresponds to the causal functional of interest, and constructs a final DR estimator by combining these elements with inverse probability weighting. The resulting estimators are shown to be root-n consistent, asymptotically normal, and semiparametrically efficient for unconditional effects, provided either the propensity score or the generative model is correctly specified. For conditional effects, we employ deep neural networks, establishing minimax-optimal convergence rates that adapt to low intrinsic data structures. Simulations confirm the double robustness and finite-sample performance of the proposed methods. This work provides a robust and flexible tool for distributional and heterogeneous causal inference in observational studies, where model misspecification is a persistent concern.

Peng Ding, University of California, Berkeley, United States

Estimating treatment effects with competing intercurrent events in randomized controlled trials

The analysis of randomized controlled trials is often complicated by intercurrent events (IEs) – events that occur after treatment initiation and affect either the interpretation or existence of outcome measurements. Examples include treatment discontinuation or the use of additional medications. In two recent clinical trials for systemic lupus erythematosus with complications of IEs, we classify the IEs into two broad categories: effect-informative (e.g., treatment discontinuation due to adverse events or lack of efficacy) and effect-uninformative (e.g., treatment discontinuation due to external factors such as pandemics or relocation). To define a clinically meaningful estimand, we adopt tailored strategies for each category of IEs. For effect-informative IEs, which are often informative about a patient's outcome, we use the composite variable strategy that assigns an outcome value indicative of treatment failure. For effect-uninformative IEs, we apply the hypothetical

strategy, assuming their timing is conditionally independent of the outcome given treatment and baseline covariates, and hypothesizing a scenario in which such events do not occur. A central yet previously overlooked challenge is the presence of competing IEs, where the first IE censors all subsequent ones. Despite its ubiquity in practice, this issue has not been explicitly recognized or addressed in previous data analyses due to the lack of rigorous statistical methodology. In this paper, we propose a principled framework to formulate the estimand, establish its nonparametric identification and semiparametric estimation theory, and introduce weighting, outcome regression, and doubly robust estimators. We apply our methods to analyze the two systemic lupus erythematosus trials, demonstrating the robustness and practical utility of the proposed framework.

DL12 — Hierarchical Poisson Species Sampling Models

June 13 (Saturday) 10:50–12:30 **Room:** LT1B Program

Lancelot James, The Hong Kong University of Science and Technology, Hong Kong

Poisson Hierarchical Indian Buffet Processes –with indications for Microbiome Models and other possibilities

We describe Poisson Hierarchical Indian Buffet Processes, designed for complex random sparse count species sampling models that facilitate information sharing across and within groups in various contexts. This model accommodates a potentially infinite number of species (taxa) and unknown parameters, allowing us to learn as more data is gathered within a Bayesian machine learning framework. We focus on the challenging context of microbiome analysis and related ecological species sampling models to address existing gaps in modelling capabilities.

Our model offers a generative process for these phenomena, providing flexible sparse multivariate count models that account for overdispersion while also modelling latent OTU counts. We present tractable methods for sampling and posterior analysis in this complex setting, introducing novel parameters that reflect species abundance as well as alpha and beta diversity. We also provide indications for novel approaches to the formidable problem of unseen entities in future samples. If time permits we may briefly discuss its broader capabilities connected to concepts of coagulation and fragmentation duality arising in probability theory.

Yi Li, University of Michigan, United States (Discussant)

Discussant

Discussant

Edwin Fong, The University of Hong Kong, Hong Kong

Predictive inference for grouped data

Bayesian hierarchical modelling is ubiquitous in many fields, but all methods require the use of Markov Chain Monte Carlo (MCMC) for posterior computation, which is computationally expensive and can be unreliable in hierarchical settings due to the complex posterior geometry. This talk discusses some potential directions for extending the martingale posterior (MP) framework, where one replaces the Bayesian prior and likelihood with a direct sequence of predictive densities, to the hierarchical setting. This involves the construction of predictive updating rules which allow for sharing between multiple groups of data. A key practical advantage is the avoidance of MCMC for posterior computation, instead relying on an imputation scheme known as predictive resampling.

IP06 — Recent Advances in Causal Inference

June 13 (Saturday) 10:50–12:30 **Room:** 209A Program

Ben Hansen, University of Michigan, United States

Design-based Hájek estimation after allocation by cluster and within blocks

Stochastic allocation to intervention conditions is essential for causal inference, but practical constraints often require assigning participants in clusters. They may be stratified pre-assignment, either of necessity or to reduce differences between treatment and control groups; but combining clustered assignment with blocking into pairs, triples, or other fine strata makes otherwise equivalent estimators perform quite differently. The two-way ANOVA with block effects can be inconsistent, as can another popular, seemingly innocuous estimator. In contrast, Hájek estimation remains broadly consistent for sample average treatment effects, but until recently lacked a design-based standard error that is applicable with clusters and fine strata. The talk justifies this assessment of current practice and outlines proposed solutions.

Yuhao Wang, Tsinghua University, China

Asymptotic theory of the best-choice rerandomization using the Mahalanobis distance

Rerandomization, a design that utilizes pretreatment covariates and improves their balance between different treatment groups, has received attention recently in both theory and practice. From a survey by Bruhn and McKenzie (2009), there are at least two types of rerandomization that are used in practice: the first rerandomizes the treatment assignment until covariate imbalance is below a prespecified threshold; the second randomizes the treatment assignment multiple times and chooses the one with the best covariate balance. In this paper we will consider the second type of rerandomization, namely the best-choice rerandomization, whose theory and inference are still lacking in the literature. In particular, we will focus on the best-choice rerandomization that uses the Mahalanobis distance to measure covariate imbalance, which is one of the most commonly used imbalance measure for multivariate covariates and is invariant to affine transformations of covariates. We will study the large-sample repeatedly sampling properties of the best-choice rerandomization, allowing both the number of covariates and the number of tried complete randomizations to increase with the sample size. We show that the asymptotic distribution of the difference-in-means estimator is more concentrated around the true average treatment effect under rerandomization than under the complete randomization, and propose large-sample accurate confidence intervals for rerandomization that are shorter than that for the completely randomized experiment. We further demonstrate that, with moderate number of covariates and with the number of tried randomizations increasing polynomially with the sample size, the best-choice rerandomization can achieve the ideally optimal precision that one can expect even with perfectly balanced covariates. The developed theory and methods for rerandomization are also illustrated using real field experiments.

Zhichao Jiang, Sun Yat-sen University, China

Principled analysis of crossover designs: causal effects, efficient estimation, and robust inference

Crossover designs randomly assign each unit to receive a sequence of treatments. By comparing outcomes within the same unit, these designs can effectively eliminate between-unit variation and facilitate the identification of both instantaneous effects of current treatments and carryover effects from past treatments. They are widely used in traditional biomedical studies and are increasingly adopted in modern digital platforms. However, standard analyses of crossover designs often rely on strong parametric models, making inference vulnerable to model misspecification. This paper adopts a design-based framework to analyze general crossover designs. We make two main contributions. First, we use potential outcomes to formally define the causal estimands and assumptions on the data-generating process. For any given type of crossover design and assumptions on potential outcomes, we outline a procedure for identification and estimation, emphasizing the central role of the treatment assignment mechanism in design-based inference. Second, we unify the analysis of crossover designs using least squares, with restrictions on the coefficients and weights on the units. Based on the theory, we recommend the specification of the regression function, weighting scheme, and coefficient restrictions to assess identifiability, construct efficient estimators, and estimate variances in a unified fashion. Crucially, the least squares procedure is simple to implement, and yields not only consistent and efficient point estimates but also valid variance estimates even when the working regression model is misspecified.

Xinran Li, The University of Chicago, United States

Randomization Inference with Sample Attrition

Randomization inference is a widely-used and appealing approach for analyzing treatment effects in randomized experiments, as it is finite-sample valid and does not require any distributional assumptions. However, naive application of randomization inference may suffer from severe size distortion in the presence of sample attrition, where outcome data are missing for some units. In this paper, we propose new, computationally efficient methods for randomization inference that remain valid under a broad class of potentially informative missingness mechanisms, allowing a unit's missingness to depend on its (unobserved) potential outcomes. Specifically, we construct valid p-values for testing both sharp and bounded null hypotheses on treatment effects via a worst-case consideration of the classical Fisher randomization test. Leveraging distribution-free test statistics, these worst-case p-values admit closed-form solutions. Importantly, by incorporating both potential outcomes and potential missingness indicators into the test statistic, our approach connects to a range of partial identification bounds in the literature, which in some sense suggests the sharpness of our tests. Moreover, our methods can exploit structural assumptions such as monotone missingness, which are commonly adopted in applications due to their plausibility and ability to substantially improve inferential power. We illustrate the proposed methods through both simulation studies and an empirical application. An R package implementing the proposed methods is publicly available.

IP09 — Recent Developments on Generative Models

June 13 (Saturday) 10:50–12:30 **Room:** 209B Program

Jian Huang, The Hong Kong Polytechnic University, Hong Kong

TBD

TBD

Xiaotong Shen, University of Minnesota, United States

Manifold-Aligned Generative Transport

High-dimensional generative modeling requires a delicate balance between support fidelity—keeping generated data precisely on the data manifold—and sampling efficiency. While diffusion models capture manifold structures well, their iterative denoising steps are slow and can leak off-support. Conversely, normalizing flows offer rapid, one-pass sampling but are limited by strict invertibility constraints. In this talk, we introduce MAGT (Manifold-Aligned Generative Transport), a novel flow-like generator that achieves the best of both worlds. MAGT learns a one-shot, manifold-aligned transport from a low-dimensional base distribution directly to the data space. By training at a fixed, numerically stable Gaussian smoothing level and approximating the score via latent anchor points, we create a tractable objective without sacrificing speed. We will demonstrate how MAGT samples in a single forward pass, tightly concentrates probability on the learned support, and enables principled likelihood evaluation. We will also share theoretical Wasserstein bounds and empirical results showing significant improvements in both fidelity and sampling speed over standard diffusion models.

This is joint work with Xinyu Tian at the University of Minnesota.

Xiaoyue Niu, The Pennsylvania State University, United States

A multilayer network model for aggregated relational data

The Network Scale-Up Method (NSUM) is a vital tool for estimating the sizes of hard-to-reach populations using Aggregated Relational Data (ARD). Recent advancements in survey design increasingly collect ARD across multiple definitions of tie strength, yielding a complex multilayer network structure. Existing statistical frameworks analyze these layers independently, sacrificing valuable shared information. We propose a latent space approach for the Multilayer ARD that enables principled joint estimation. We establish rigorous identifiability conditions for the shared latent space and mathematically prove that the joint estimator achieves asymptotic efficiency gains and finite sample bias reductions by "borrowing strength" across layers. We apply the framework to real-world multilayer NSUM survey data, demonstrating its practical efficacy in yielding more robust and precise subpopulation estimates.

IP08 — Recent Developments on Network Analysis and Biostatistics

June 13 (Saturday) 10:50–12:30 **Room:** 203 Program

Ji Zhu, University of Michigan, United States

Statistical Inference for Latent Space Models of Network Data with Edge Covariates

Latent space models (LSMs) provide a powerful framework for analyzing network data by embedding nodes in a latent space. Incorporating covariate information via edge covariates offers an important generalization that strengthens both the interpretability and practical utility of the model. However, we show that coefficient estimates for edge covariate effects obtained through maximum likelihood estimation exhibit asymptotic bias due to high-order geometric effects and errors in latent variable estimation. To address this issue, we propose a plug-in bias-correction estimator that enables asymptotically valid and unbiased statistical inference for the effects of edge covariates. We establish theoretical guaranties, including consistency and asymptotic normality, under various network structures. Extensive simulations and real-world data examples demonstrate that our method effectively reduces estimation bias and improves the accuracy of inference. Our findings contribute to the statistical methodology of LSMs by providing a principled framework for unbiased parameter estimation in network models with edge covariates.

Haoran Xue, City University of Hong Kong, Hong Kong

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

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

Lu Tian, Stanford University, United States

An Honest Cross-Validation Estimator for Prediction Performance

Cross-validation is a standard tool for obtaining an honest assessment of the performance of a prediction model. The commonly used version repeatedly splits data, trains the prediction model on the training set, evaluates the model performance on the test set, and averages the model performance across different data splits. A well-known criticism is that such cross-validation procedure does not directly estimate the performance of the particular model recommended for future use. In this paper, we propose a new method to estimate the performance of a model trained on a specific (random) training set. A naive estimator of the prediction performance can be obtained by applying the model to a disjoint testing set. Surprisingly, cross-validation estimators computed from other random splits can be used to improve this naive estimator within a random-effects model framework. We develop two estimators—a hierarchical Bayesian estimator and an empirical Bayes estimator—that perform similarly to or better than both the conventional cross-validation estimator and the naive single-split estimator. Simulations and a real-data example demonstrate the superior performance of the proposed method.

Annie Qu, University of California, Irvine, United States

Dynamic Topic Modeling with a Higher-Order Hypergraphical Representation

Dynamic topic modeling is widely used to analyze evolving trends in scientific literature, medical records, and social media. Traditional topic models represent each topic through a single probability vector on the multinomial simplex and implicitly couple word occurrence and repetition within one probabilistic mechanism. However, this formulation restricts the dependence structure among words and overlooks informative higher-order interactions, particularly in dynamic corpora with overlapping semantics. To address these limitations, we introduce a hypergraphical representation of text where each document is modeled as a hyperedge connecting all co-occurring words, with repetition intensities encoded as node weights. This representation naturally separates word occurrence from repetition and induces a novel hypergraph-based multinomial distribution with a nonlinear normalization depending on the observed word set of each document. Building on this likelihood, we develop a dynamic topic modeling framework via structured low-rank factorizations with explicit temporal regularization on topic–word profiles. Moreover, in theory, we establish local convergence guarantees and derive non-asymptotic error bounds despite the intrinsic nonconvexity of bilinear factorization and document-specific nonlinear normalization. Numerical experiments on synthetic data and an application to the International Conference on Learning Representations (ICLR) corpus demonstrate consistent improvements over existing multinomial-based topic models.

IP61 — Advances in the Design and Analysis of Clinical Trials

June 13 (Saturday) 10:50–12:30 **Room:** 201 Program

Wei Zhang, Chinese Academy of Sciences, China

Optimal treatment allocations accounting for population differences

The treatment allocation mechanism in a randomized clinical trial can be optimized by maximizing the nonparametric efficiency bound for a specific measure of treatment effect. Optimal treatment allocations which may or may not depend on baseline covariates have been derived for a variety of effect measures focusing on the trial population, the patient population represented by the trial participants. Frequently, clinical trial data are used to estimate treatment effects in a target population that is related to but different from the trial population. This article provides optimal treatment allocations that account for the impact of such population differences. We consider three cases with different data configurations: transportation, generalization, and post-stratification. Our results indicate that, for general effect measures, optimal treatment allocations may depend on the covariate distribution in the target population but not on the configuration of data or information that describes the target covariate distribution. For estimating average treatment effects, there is a unique covariate-dependent allocation that achieves maximal efficiency regardless of the target covariate distribution and the associated data configuration.

Yu Jun, Beijing Institute of Technology, China

Minimum free energy driven randomized allocation to improve covariate balance

The tension between randomization for robustness and covariate balancing for inferential efficiency remains a central debate in clinical trial design. Randomized allocation provides diverse allocations to enhance the robustness of the treatment effect estimation. The covariate balance strategies maximize the inference efficiency under some hypothesized data-generating models. Drawing an analogy to the second law of thermodynamics—where system equilibrium minimizes Helmholtz free energy (balancing internal energy and disorder)—we propose a novel randomized allocation framework to navigate this trade-off. Efficient one-shot and sequential allocation algorithms are designed to harmonize randomness with covariate balance. Finite-sample theoretical guarantees for these procedures are established. Simulation studies and a real-world application confirm the theoretical advantages of the proposed design strategy, demonstrating superiority in treatment effect estimation.

Yang Liu, Renmin University of China, China

The Impact of Unobserved Covariates on Covariate-Adaptive Randomized Experiments

Covariate-adaptive randomization (CAR) procedures are widely used in randomized control studies to achieve balance in observed covariates, thereby improving statistical efficiency and validity. Despite their success, two fundamental questions remain affecting the application of these methods: (1) Does balancing observed covariates also improve balance in unobserved covariates? (2) Is such balance sufficient to ensure valid inference of treatment effect? In this talk, I will present our recent works addressing these questions in the context of CAR with discrete covariates. Regarding the first question, I will introduce a theoretical framework for evaluating unobserved covariates imbalance. Our findings show that balancing observed covariates often leads to improved balance in unobserved ones, providing theoretical support for the advantage of CAR over complete randomization. For the second question, I will examine the inference of treatment effect under a linear outcome model. We show that when unobserved covariates interact with treatment, the treatment effect may become non-identifiable, leading to inconsistent estimation. Even when no interaction exists, standard model-based inference procedures may exhibit reduced Type I error rates. I will discuss robust adjustment methods that restore valid inference under CAR. These findings offer both theoretical insight and practical guidance for the better application of CAR.

Sai Li, Tsinghua University, China

Personalizing black-box models for nonparametric regression with minimax optimality

Recent advances in large-scale models, including deep neural networks and large language models, have substantially improved performance across a wide range of learning tasks. The widespread availability of such pre-trained models creates new opportunities for data-efficient statistical learning, provided they can be effectively integrated into downstream tasks. Motivated by this setting, we study few-shot personalization, where a pre-trained black-box model is adapted to a target domain using a limited number of samples. We develop a theoretical framework for few-shot personalization in nonparametric regression and propose algorithms that can incorporate a black-box pre-trained model into the regression procedure. We establish the minimax optimal rate for the personalization problem and show that the proposed method attains this rate. Our results quantify the statistical benefits of leveraging pre-trained models under sample scarcity and provide robustness guarantees when the pre-trained model is not informative. We demonstrate the finite-sample performance of the methods through simulations and an application to the California housing dataset with several pre-trained models.

IP25 — Advances in Statistical Methods for Complex Data and Clinical Studies

June 13 (Saturday) 10:50–12:30 **Room:** 202 Program

Lijian Yang, Tsinghua University, China

Simultaneous inference for finite distribution functions of stochastic processes

Simultaneous confidence regions (SCRs) are proposed for finite distribution functions of stochastic processes with any asymptotically correct confidence level. For discrete time processes, SCRs are derived from empirical cumulative distribution function (ECDF) and kernel distribution estimator (KDE) respectively, under mild assumptions of continuity and H\"{o}lder continuity respectively. For continuous-time processes, SCRs are derived from two-step KDE where an initial spline regression step estimates all continuous sample trajectories. Continuous-time processes with nondifferentiable sample paths or heavy tail finite distribution functions are both included in the theoretical development. Also formulated are Lower SCRs (LSCRs) and upper SCRs (USCRs) to allow for one-sided inference. Rigorous theory is developed for the continuity and strict monotonicity for all extreme value distributions of the limiting multivariate Gaussian process, justifying the use of SCRs in hypothesis testing by asymptotically exact p-values. Simulation examples provide ample evidence of the SCRs validity, and hypothesis testing on finite distribution functions of an electroencephalogram (EEG) data shows clearly the appropriateness of Gaussianity assumption for the EEG process.

Lucy Xia, The Hong Kong University of Science and Technology, Hong Kong

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

Recent clinical studies increasingly involve a long factor with many levels, e.g., investigation sites, resulting in a large

number of strata that must be accounted for either through design or subsequent analysis. This complication has raised concerns by the U.S. Food and Drug Administration regarding the adequacy of standard statistical methods, whose performance may deteriorate, and their properties become unclear when the number of strata is relatively large. In this work, we offer a first-time rigorous solution by employing mixed-effect models in covariate-adaptive randomized experiments. We show that the mixed-effect estimate achieves lower variance in treatment effect estimation than its fixed-effect counterpart in the presence of the long factor with many levels. This variance reduction is attributable to marginal imbalances induced by the randomization procedure, suggesting that designs promoting finer covariate balance lead to more efficient inference and increased statistical power. Furthermore, we demonstrate that, as the sample size grows, the mixed- and fixed-effect estimators become asymptotically equivalent. Our theoretical findings are validated through simulations and a clinical trial case study, and provide new insights on the design and analysis of clinical trials in the presence of a large number of strata.

Shunan Yao, Hong Kong Baptist University, Hong Kong

U-processes and their application in mean estimation

This talk presents new results on robust multivariate location estimation built on U-statistical methodology. First, we derive a deviation a Bousquet-type inequality for U-processes with near-optimal constants. Second, using the derived technical results, we introduce a new location estimator based on a U-statistic version of Tukey's median. We show that the proposed estimator is asymptotically normal and achieves efficient asymptotic covariance, in contrast to the original Tukey's median.

Yifan Chen, Hong Kong Baptist University, Hong Kong

A Recipe for Causal Graph Regression: Confounding Effects Revisited

AI systems learning from complex networks (graphs) often struggle when faced with unfamiliar data. While "causal learning" helps these systems focus on true causes, making them more robust, it has primarily been applied to categorizing items rather than predicting specific numbers—a tougher challenge for AI on graphs. Our research introduces a new method for applying causal learning to predict numbers on graphs. This approach addresses misleading information, known as "confounders," by adapting established techniques. Furthermore, we utilize "contrastive learning." This enables us to successfully extend causal methods, which were previously specific to classification (sorting items), to the task of regression (predicting numerical values). The result is AI that can make more reliable numerical predictions on graphs, even as data changes. This is crucial for developing trustworthy AI in fields like materials science or economics, where data is inherently complex and ever-evolving.

IP45 — Applied Probability and Financial Mathematics

June 13 (Saturday) 10:50–12:30 **Room:** 214 Program

Ju-Yi Yen, University of Cincinnati, United States

An arbitrage driven price dynamics of Automated Market Makers in the presence of fees

Under a reference market price following a geometric Brownian motion, we present a model for price dynamics in the Automated Market Makers (AMM) setting. The AMM price is constrained within bounds determined by constant multiples of this reference price. By employing local time and excursion theory, we derive several analytical results, including a time-changed representation of the AMM price process and its asymptotic behavior.

Chi Seng Pun, Nanyang Technological University, Singapore

Pairs Trading with Frictions: Transaction Costs, Market Impact, and Optimal Strategies

Pairs trading traditionally assumes frictionless markets, yet real-world execution is shaped by transaction costs, market impact, and finite investment horizons. In this talk, we present a unified framework for understanding how different forms of trading frictions alter the optimal behavior of statistical arbitrage strategies. We first study pairs trading with proportional transaction costs under a finite horizon. The problem leads to a high-dimensional singular stochastic control formulation,

whose value function satisfies a Hamilton–Jacobi–Bellman variational inequality with endogenous free boundaries. Using asymptotic expansion and homogenization techniques, we derive tractable characterizations of the no-trade region and construct a nearly optimal strategy that captures the leading-order welfare impact of transaction costs.

We then examine a complementary setting in which trading incurs market impact costs. Here, the controls are continuous trading rates, and the optimal strategy is explicitly characterized by solving a nonlinear HJB equation under CARA utility. This yields closed-form expressions for optimal trading intensity, highlighting the interplay between mean reversion, impact parameters, and risk preferences.

Together, these two models offer a comprehensive view of how different frictions reshape classical pairs trading rules. The results provide both theoretical insights and practical guidance for implementing pairs trading in realistic markets.

Chuan-Hsiang Han, National Tsing Hua University, Taiwan

Accelerating Deep Learning by Efficient Importance Sampling for CVaR Estimation

This talk presents an accelerated deep learning framework for Conditional Value-at-Risk (CVaR) estimation. We parameterize the likelihood ratio in the dual representation with a multilayer perceptron and integrate efficient importance sampling (EIS) into the stochastic gradient scheme on GPUs. EIS achieves asymptotically minimal variance, significantly reducing the number of samples required and lowering computational cost. Numerical simulations demonstrate faster training and improved stability, highlighting the synergy between asymptotically optimal Monte Carlo methods and deep learning for scalable risk management.

Li-Hsien Sun, National Central University, Taiwan

Partial Information in a Mean-Variance Portfolio Selection Game

We consider finitely many investors who perform mean-variance portfolio selection under a relative performance criterion. That is, each investor is concerned about not only her terminal wealth, but how it compares to the average terminal wealth of all investors (i.e., the mean field). We derive such a Nash equilibrium explicitly in the idealized case of full information (i.e., the dynamics of the underlying stock is perfectly known), and semi-explicitly in the realistic case of partial information (i.e., the stock evolution is observed, but the expected return of the stock is not precisely known). The formula under partial information involves an additional state process that serves to filter the true state of the expected return. We comment on the effect of partial information through numerical analysis. Observe that partial information alone can reduce investor's wealth significantly, thereby causing or aggravating systemic risk.

IP47 — Design and Analysis of Modern Experiments

June 13 (Saturday) 10:50–12:30 **Room:** 215 Program

Ming-Chung Chang, Academia Sinica, Taiwan

Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs

Multi-stratum factorial designs, such as block designs and row–column designs, are widely used for screening treatment factors in experiments with complex experimental-unit structures arising from multiple sources of variability. In this presentation, I will introduce a unified, model-free approach, termed orthogonalized moment aberration, for comparing similarities among level combinations of treatment factors assigned to heterogeneous experimental units. The proposed approach evaluates the rows of design matrices through kernel functions, rather than the columns, enabling the assessment of a broad class of mixed-level regular and nonregular factorial designs under heterogeneous experimental-unit structures known as partially relaxed orthogonal block structures. This framework is highly flexible: different choices of kernel functions allow adaptation to various experimental scenarios, with certain choices recovering well-known minimum aberration criteria from the literature. Although model-free in nature, the proposed method admits rigorous justification via linear mixed-effects models and Gaussian process models. Theoretical results and numerical examples demonstrate that this approach can generate multi-stratum factorial designs with high Bayesian D-efficiencies.

Jing-Wen Huang, Academia Sinica, Taiwan

Cyclic Order-of-addition experiments

Order-of-addition experiments investigate how the outcome of a process depends on the sequence in which a fixed set of components is introduced. Unlike standard factorial designs, the same components are present in every run, but their ordering induces potentially complex and noncommutative effects on the response.

Existing work on order-of-addition experiments typically focuses on sequences that are applied once or terminate after a finite number of steps. However, in many practical settings, the same ordered sequence may be applied repeatedly in a cyclic manner, giving rise to data structures that are not adequately accommodated by current order-of-addition models.

In this work, we propose a distance-based linear model to account for such repeated order structures, along with a cyclic version of complete consecutive order-pairing design tailored to the proposed framework. Simulation studies, which are related to the coordinate descent algorithm and the traveling salesman problem, further demonstrate the effectiveness of proposed approach.

William Li, Shanghai Advanced Institute of Finance, China

Robust Integer-Valued Designs for Non-Linear Models: An Algorithmic Approach with Efficiency Guarantee

Robust experimental design under model misspecification has been studied extensively using continuous designs and, in a more limited literature, integer-valued designs for linear and general linear models. In this paper, we study integer-valued robust designs for nonlinear regression models under an average mean squared error criterion. The robust loss decomposes into variance and bias components, whose competing effects on design configuration have been well recognized in earlier work but have typically been handled by optimizing the combined criterion directly using simulated annealing. We show that, when designs are required to be integer-valued, optimization of the bias component alone reduces to a bounded subset-selection problem. Leveraging on recent advances in algorithms for bounded designs, we develop an efficient algorithmic approach for constructing integer-valued robust designs with explicit efficiency bounds. Although the methodology is developed for nonlinear models, the proposed approach applies more broadly and can be used to strengthen existing results for linear and related models. Numerical examples demonstrate improved efficiency, stability, and scalability relative to simulated-annealing-based methods.

DL05 — To be confirmed

June 13 (Saturday) 13:30–15:10 **Room:** LT1A Program

Jing Lei, Carnegie Mellon University, United States

Evaluating Black-Box Classifiers via Stable Adaptive Two-Sample Inference

Evaluating the quality of black-box classifiers is a fundamental problem in statistics and machine learning. Building upon a previous framework which evaluates classifiers using a tolerant goodness-of-fit testing, our approach reduces the evaluation task to a two-sample conditional distribution testing problem by generating an auxiliary sample from the fitted classifier and training a distinguisher" to differentiate between the true and generated samples. The distinguisher' s ability, quantified through a rank-sum statistic, measures the discrepancy between the estimated and true conditional distributions. Using techniques from cross-validation central limit theorems, we derive an asymptotically rigorous test under suitable stability conditions of the distinguisher. The empirical performance of our method are demonstrated using both numerical simulations and real-world datasets.

David Rügamer, University of Munich, Germany

Smooth Optimization for Sparse Learning via Overparameterization

Sparse regularization is essential for interpretable and efficient machine learning, but classical penalties are often non-smooth and difficult to combine with the gradient-based training procedures prevalent in deep learning. This talk discusses

how overparameterization can address this challenge by replacing non-smooth objectives in the original parameters with smooth optimization in an expanded parameter space. The resulting methods recover classical and structured sparsity penalties while remaining compatible with modern training pipelines. Theoretical results characterize the induced penalties and connect the resulting optima and dynamics to sparse solutions.

DL01 — Structured Nonparametrics

June 13 (Saturday) 13:30–15:10 **Room:** LT1B Program

Enno Mammen, Heidelberg University, Germany

Additive nonparametric regression: interaction, high dimension and beyond

Structured nonparametric regression is to estimate the regression function within a class of functions that are expressed in terms of component functions defined on lower dimensional spaces. For many models it has been pointed out that one may estimate the regression function at the same accuracy as one may achieve in estimating lower dimensional functions regardless of the actual dimension of the covariate vector. A leading and very illustrative example are additive nonparametric regression models. In this talk we discuss some recent developments on models related to additive models. This includes additive models with interaction terms for which it is difficult to carry over L_2 results to an L_∞ framework which is needed to develop asymptotic theory for functional estimators. We discuss models where the number of components converge to infinity and we consider additive models with variables in Riemannian Hilbert manifolds. In our presentation we concentrate on methods using smooth backfitting besides a short excursion to random additive trees.

Wolfgang Polonik, University of California, Davis, United States

On the structure of the Euler characteristic of a VR-filtration

The Euler characteristic of a Vietoris-Rips filtration constructed over a point cloud sampled from a manifold plays an important role in geometric and topological data analysis. In this talk, we will introduce and discuss structural aspects of this quantity and the information captured by it. Moreover, we will show how to separate information on the geometry of the manifold from information on the sampling distribution, and discuss implications for statistical inference. This is joint work with Johannes Krebs, Catholic University of Eichstätt-Ingolstadt.

Changwon Choi, Seoul National University, South Korea

Additive Fréchet regression for random objects

Regression analysis for complex data taking values in a general metric space has gained increasing attention in recent years, particularly in the context of Fréchet regression with Euclidean predictors. However, local Fréchet regression, while more flexible than global Fréchet regression, suffers from the curse of dimensionality for the case of multivariate predictors. To address this issue while maintaining model flexibility, we introduce a framework of additive structured nonparametric regression when responses are situated in a general metric space. Overcoming the challenge of the lack of vector space structure in general metric spaces where the responses reside requires a novel approach that incorporates the additive structure via projection operators. This leads to a unified framework of additive regression for a wide range of response types, including one-dimensional distributions in Wasserstein space, network data represented by graph Laplacians and spherical data equipped with geodesic distances, among others. We establish pointwise and uniform consistency, including convergence rates, for the proposed additive Fréchet regression estimators, using smooth backfitting. The practical utility of this additive model is demonstrated through applications to brain connectivity network analysis using resting-state fMRI data and distributional physical activity data.

IP04 — Statistical Opportunities in Deep and Reinforcement Learning

June 13 (Saturday) 13:30–15:10 **Room:** 209A Program

Guohao Shen, The Hong Kong Polytechnic University, Hong Kong

Symmetries in Deep Neural Networks and Implications to Learning

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

Faming Liang, Purdue University, United States

Uncertainty Quantification for Physics-Informed Neural Networks with Extended Fiducial Inference

Uncertainty quantification (UQ) in scientific machine learning is increasingly critical as neural networks are widely adopted to tackle complex

problems across diverse scientific disciplines. For physics-informed neural networks (PINNs), a prominent model in scientific machine learning, uncertainty is typically quantified using Bayesian or dropout methods. However, both approaches suffer from a fundamental limitation: the prior distribution or dropout rate required to construct honest confidence sets cannot be determined without additional information. In this work, we propose a novel method within the framework of extended fiducial inference (EFI) to provide rigorous uncertainty quantification for PINNs. The proposed method leverages a narrow-neck hyper-network to learn the parameters of the PINN and quantify their uncertainty based on imputed random errors in the observations. This approach overcomes the limitations of Bayesian and dropout methods, enabling the construction of honest confidence sets based solely on observed data.

This advancement represents a significant breakthrough for PINNs, greatly enhancing their reliability, interpretability, and applicability to real-world scientific and engineering challenges. Moreover, it establishes a new theoretical framework for EFI, extending its application to large-scale models, eliminating the need for sparse hyper-networks, and significantly improving the automaticity and robustness of statistical inference. This is joint work with Frank Shih and Zhenghao Jiang.

Chengchun Shi, The London School of Economics and Political Science, United Kingdom

Demystifying LLM Reasoning through the Lens of U-Statistics

Group relative policy optimization (GRPO), a core methodological component of DeepSeekMath and DeepSeek-R1, has emerged as a cornerstone for scaling reasoning capabilities of large

language models. Despite its widespread adoption and the proliferation of follow-up works, the theoretical properties of GRPO remain less studied. This talk provides a unified framework to

understand GRPO through the lens of classical U-statistics. We demonstrate that the GRPO policy gradient is inherently a U-statistic, allowing us to characterize its mean squared error

(MSE), derive the finite-sample error bound and asymptotic distribution of the suboptimality gap for its learned policy. Our findings reveal that GRPO is asymptotically equivalent to an oracle policy gradient algorithm −one with access to a value function that quantifies the goodness of its learning policy at each training iteration −and achieves asymptotically optimal performance

within a broad class of policy gradient algorithms. Furthermore, we establish a universal scaling law that offers principled guidance for selecting the optimal group size. Empirical experiments

further validate our theoretical findings, demonstrating that the optimal group size is universal, and verify the oracle property of GRPO.

Hongtu Zhu, The University of North Carolina at Chapel Hill, United States

Causal deepsets for off-policy evaluation under spatial or spatio-temporal interferences

Off-policy evaluation (OPE) is widely applied in sectors such as pharmaceuticals and e-commerce to evaluate the efficacy of novel products or policies from offline datasets. This paper introduces a causal deepset framework that relaxes several key structural assumptions, primarily the mean-field assumption, prevalent in existing OPE methodologies that handle spatio-temporal interference. These traditional assumptions frequently prove inadequate in real-world settings, thereby restricting the capability of current OPE methods to effectively address complex interference effects. In response, we advocate for the implementation of the permutation invariance (PI) assumption. This innovative approach enables the data-driven, adaptive learning of the mean-field function, offering a more flexible estimation method beyond conventional averaging. Furthermore, we present novel algorithms that incorporate the PI assumption into OPE and thoroughly examine their theoretical foundations. Our numerical analyses demonstrate that this novel approach yields significantly more precise estimations than existing baseline algorithms, thereby substantially improving the practical applicability and effectiveness of OPE methodologies. A Python implementation of our proposed method is available at https://github.com/BIG-S2/Causal-Deepsets.

IP70 — Recent Development on AI and Biostatistics

June 13 (Saturday) 13:30–15:10 **Room:** 209B Program

Jie Ding, University of Minnesota, United States

The Future of AI Scientists: Emerging Directions and Fundamental Challenges

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.

Ji Yuan, The University of Chicago, United States

From research idea to peer-reviewed draft: an agentic pipeline for statistical methodology research with adversarial revision

We describe an agentic system that takes a single research idea through scoping, methodology design, validation, manuscript writing, and adversarial peer review, producing a submission-ready statistical paper with no human input between phases. ARMS (Autonomous Research Manuscript Skills) is an 11-skill pipeline. Each phase runs as a separate agent invocation and communicates with the next through files on disk, which prevents information degradation from agent-to-agent summarisation. Phase 2 validates the proposed method against pre-specified comparators and success criteria before any manuscript text is written; the pipeline can return KILL or NO-GO when the method fails, treating honest failure as a first-class outcome. We propose a Critic-Revisor: two parallel Opus reviewers per round adopt differentiated angles, one as methodologist and one as applied and contextual reader, each issuing a capped, ranked critique with quoted passages and concrete fixes. Reviewers may web-search within a tiered per-round budget so claims and citations are externally verified rather than hallucinated. A separate Codex agent applies the combined critique; latexmk and latexdiff produce a clean PDF and a colour-tracked diff each round. The loop ends when both reviewers accept, the build fails, or the round budget is exhausted. We illustrate on Bayesian clinical-trial methodology, including a fully autonomous design in independent grading. We discuss what we have learned: the persistent gap between self-grading and independent grading, the limits of agentic frontier-pushing, and the implications for the practice and reviewing of statistical research.

Joan J. Ren, University of Maryland, United States

Empirical Likelihood Based Multivariate Distribution Estimator for Various Types of Censored Data and Its Application in Censored Causal Inference

Speaker:

Joan Jian-Jian Ren

Statistics Program, Department of Mathematics,

University of Maryland - College Park

Abstract: The analysis of censored multivariate data is of great importance in medical research, reliability studies, epidemiology, social and behavior science studies, etc. But in practice, we often encounter censored multivariate data which do not fit the existing parametric or semiparametric model assumptions, or often no existing goodness-of-fit tests for model checking. In current statistical literature, there is no existing multivariate distribution estimator for distribution function F_0(t, z) of (T, Z) based on various types of censored multivariate data for survival time T and covariate vector $Z \in R^p$ containing continuous components with p > 1. In the context of causal inference, the conditional average treatment effect (CATE) can play an important role to study the impact of covariate Z on survival time T , which is often subject to censoring in practice. So far, most existing works on CATE are based on parametric or semiparametric model assumptions, and few deals with censored multivariate data. This article constructs empirical likelihood based multivariate distribution estimator (ELB-MDE) for F_0(t, z) with various types of censored multivariate data, such as right censoring, double censoring, interval censoring, partly interval-censoring, etc., which is completely nonparametric, innovative and the first unified framework that leads to the estimation of CATE based on various types of censored multivariate data. The asymptotic properties of our ELB-MDE and CATE estimator are established, and some simulation results as well as HIV data analysis results are presented.

Coauthor:

Charles Zhao

Department of Statistics & Operations Research,

University of North Carolina - Chapel Hill

IP24 — Classification, Community Detection and Inference for High Dimensional Complex Data

June 13 (Saturday) 13:30–15:10 **Room:** 203 Program

Jinchi Lv, University of Southern California, United States

HNCI: High-Dimensional Network Causal Inference

We propose a new method of high-dimensional network causal inference (HNCI) that provides both valid confidence intervals for the average direct treatment effect on the treated (ADET) and valid confidence sets for the neighborhood size affecting the interference effect. We adopt the model framework from Belloni et al. (2022), which has key advantages such as nonparametric modeling of interference effects and allowing certain types of heterogeneity in node interference neighborhood sizes. Utilizing the nested matching property of the network interference effect, we reformulate the original nonparametric model into a linear regression model where the regression coefficients, corresponding to the underlying true interference function values of nodes, exhibit a latent homogeneous structure. This formulation enables us to leverage existing literature on homogeneity pursuit to conduct valid statistical inferences with theoretical guarantees. The resulting confidence intervals for the ADET are formally justified through asymptotic normality with estimable variances. By employing the repro samples approach, we further provide the confidence set for the interference of neighborhood size with theoretical guarantees. The practical utility of the newly suggested methods is demonstrated through simulations and real data examples. This is a joint work with Rundong Ding, Wenqin Du and Yingying Fan.

Yin Xia, Fudan University, China

A Unified Framework for Large-Scale Inference of Classification: Error Rate Control and Optimality

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 algorithmic 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. Additionally, 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.

Yi-Hsin Yang, National Health Research Institutes, Taiwan

Determining lines of therapy algorithms to detect cancer progression events in electronic health records

The lines of therapy (LOTs) algorithms used in structured electronic health records (EHRs) have been widely developed as surrogates of cancer progression. However, changes in cancer therapy may also be due to intolerable toxicity. To improve the detection of disease progression, implementing large language models (LLMs) in unstructured EHRs may enhance accuracy. This study aimed to investigate the performance of detecting cancer progression by combining the LOT algorithms with LLMs in metastatic colorectal cancer (mCRC) patients.

This retrospective cohort study used the Kaohsiung Medical University Hospital Research Database (KMUHRD) to retrieve patients' medication records, pathological reports, and imaging reports from 2011 to 2020. An mCRC LOT algorithm was developed, and changes in LOT were considered a surrogate for cancer progression. Two LLMs of different scales, Google/gemma-3-27b-it and Deepseek-ai/DeepSeek-R1-Distill-Llama-70B, were employed to identify keywords related to progression or metastasis. Their results were subsequently validated using the Cancer Case Management data in KMUHRD. Model performance was evaluated using accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score.

We included 806 patients, of whom 456 (56.6%) had a history of progression. The mCRC LOT algorithm achieved an accuracy of 71.2%, a sensitivity of 66.9%, a specificity of 76.9%, a PPV of 79.0%, a NPV of 64.0%, and an F1 score of 0.724. After combining (union) the results of the LLM (Google/gemma-3-27b-it with single-task instruction), the F1 score was 0.745. The sensitivities (recall) became better in the union of LOT, and LLM (70.6%~90.8%), and the specificities became better in the intersection of LOT and LLM (80.9%~96.9%).

We successfully integrated LOT algorithms with LLMs to identify progression in mCRC patients using EHRs. These algorithms and AI tools can now help conduct comparative effectiveness studies using EHRs more efficiently. Furthermore, the algorithms/ AI tools developed in this study may be extended to other cancers. The findings indicate that focusing on a single task allows the models to execute keyword analysis with greater precision. This suggests that simplifying task complexity is a viable strategy for improving LLM performance in specialized applications.

IP44 — Recent Advances in Sampling Methods

June 13 (Saturday) 13:30–15:10 **Room:** 201 Program

Zi Yang Meng, The University of Hong Kong, Hong Kong

(Artificial) intelligent Monte Carlo sampling in quantum many-body systems

I will use a few example in the modern quantum many-body computational research, to show how important to generate the (not necessarily artificial) intelligent Monte Carlo sampling in these systems, such that the correct estimation of the expectation value and error can be achieved. Often time, the simple but profound concept of important sampling in a Monte Carlo process is masked by the physical complexity of the problem at hand, but one needs to overcome these obstacles (via mathematical intuition) to reveal the true understanding.

Zijing Ou, Imperial College London, United Kingdom

Neural Flow Samplers: Improved Training and Architectures

Sampling from unnormalized densities, either for continuous or discrete variables, presents a fundamental challenge with wide-ranging applications, from posterior inference to molecular dynamics simulations and combinatorial optimisation.

Continuous or discrete flow-based neural samplers offer a promising approach, learning a velocity field for continuous variable densities (or rate matrix for a continuous-time Markov chain (CTMC) built for discrete variable densities) that satisfies key principles of marginal density evolution (e.g., the continuity equation for continuous variable case and the Kolmogorov equation for discrete variable case) to generate samples. However, this learning procedure requires accurate estimation of intractable terms linked to the computationally challenging partition function, for which existing estimators often suffer from high variance or low accuracy. To overcome this, we introduce an improved estimator for these challenging quantities, employing an improved Sequential Monte Carlo method enhanced with control variates. We further introduce specific adjustments and advances for the trained sampler tailored for continuous variable or discrete variable case. In both scenarios, our proposed Neural Flow Samplers empirically outperforms existing flow-based neural samplers on both synthetic datasets and complex targets motivated by real-world applications.

Miha Bresar, The Chinese University of Hong Kong-Shenzhen, China

TBD

TBD

Jiajun He, University of Cambridge, United Kingdom

Diffusion Model Control with Monte Carlo Methods

TBD

IP60 — Some of the Latest Advances in Multiple Testing

June 13 (Saturday) 13:30–15:10 **Room:** 202 Program

Jesse Hemerik, Erasmus University Rotterdam, Netherlands

Resampling-based multi-resolution false discovery exceedance control

MaxT is a highly popular resampling-based multiple testing procedure, which controls the Familywise Error Rate (FWER) and is powerful under dependence. This paper generalizes maxT to what we term "multi-resolution" False Discovery eXceedance (FDX) control. Basic FDX control means ensuring that the FDP —the proportion of false discoveries among all rejections —is at most γ with probability at least $1-\alpha$. Here γ and α are prespecified, small values between 0 and 1. Importantly, our method is in addition simultaneous, in the following way: the procedure outputs a single rejection threshold q, but ensures that with probability $1-\alpha$, simultaneously over all stricter thresholds, the corresponding FDPs are also below γ. In particular, for a small set of hypotheses, the FDP bound is 0, i.e., the FWER is 0. Despite these additional, simultaneous guarantees, our method has power comparable to Romano-Wolf, the most powerful non-simultaneous FDX method. Further, our method is valid under the same assumptions. Thus, this paper shows that FDX methods can often be made simultaneous almost for free. Link to paper: https://arxiv.org/abs/2509.02376

Shinjini Nandi, Montana State University, United States

Leveraging the group structure of hypotheses for more powerful multiple testing with FDR control for the filtered rejection set

Modern biological studies often involve testing large number of hypotheses organized in groups and/or in a hierarchical structure, such as a directed acyclic graph (DAG). In these studies, researchers often wish to control the false discovery rate (FDR) after 'filtering' the discoveries to obtain interpretable results. In order to address this goal, Katsevich, Sabatti, and Bogomolov (2023, Journal of the American Statistical Association, 118(541), 165-176) developed a general method, Focused BH, that guarantees FDR control for the filtered set of discoveries made by a multiple testing procedure, for a pre-specified filter, under certain assumptions. We propose improving the power of Focused BH by adapting it to grouped and hierarchical structures of hypotheses using data-dependent weights. The general method incorporating such weights is

referred to as Weighted Focused BH (WFBH). For DAG-structured hypotheses, we propose a variant of WFBH, which gains power by adapting to the DAG structure, and by leveraging the logical relationships among the hypotheses. We prove that a variant of WFBH designed for testing different group-structures, as well as its proposed variant for testing DAG-structured hypotheses, control the post-filtering FDR under certain assumptions. Through simulations, we demonstrate that the latter variant is robust to deviations from these assumptions and can be considerably more powerful than comparable methods. Finally, we elucidate its practical use by applying it to real datasets from microbiome and gene expression studies.

Jinzhou Li, National University of Singapore, Singapore

Simultaneous false discovery proportion bounds via knockoffs and closed testing

We propose new methods to obtain simultaneous false discovery proportion bounds for knockoff-based approaches. We first investigate an approach based on Janson and Su's k-familywise error rate control method and interpolation. We then generalize it by considering a collection of kvalues, and show that the bound of Katsevich and Ramdas is a special case of this method and can be uniformly improved. Next, we further generalize the method by using closed testing with a multi-weighted-sum local test statistic. This allows us to obtain a further uniform improvement and other generalizations over previous methods. We also develop an efficient shortcut for its implementation. We compare the performance of our proposed methods in simulations and apply them to a data set from the UK Biobank.

Qiuqi Wang, Georgia State University, United States

False discovery rates of refreshing monitoring procedures

Practical monitoring procedures require anytime validity. In this talk, we examine the rationality of monitoring procedures that restart whenever rejections occur. Specifically, we study an online false discovery rate (FDR) control problem of refreshing monitoring procedures based on test (super)martingales and e-processes. We obtain explicit FDR bounds with fixed rejection thresholds. Moreover, we improve refreshing monitoring procedures by dropping less informative data points with an FDR guarantee. Numerical examples of refreshing monitoring procedures will be presented, including applications to mortality modeling and financial regulatory backtests.

IP59 — Advances in Forecasting and Time Series Analysis

June 13 (Saturday) 13:30–15:10 **Room:** 214 Program

Yu Fu, Melbourne Business School, Australia

Regression Copula Process MIDAS for Macroeconomic Density Forecasting

We show how to calibrate margins of response variables in mixed-data sampling (MIDAS) regressions using a copula regression process. The copula-based approach separates marginal distributions from dependence generated by the regression structure. This allows the predictive distribution to accommodate skewness and tail behaviour while retaining the interpretability and computational efficiency from a MIDAS framework. In an expanding-window forecasting exercise using U.S. macroeconomic data, our model is evaluated for GDP growth and GDP deflator inflation at horizons of one to six months. Relative to Gaussian MIDAS benchmarks, the copula process approach delivers substantial gains for density forecasts of inflation.

Chao Wang, The University of Sydney, Australia

Combining Forecasts of Value-at-Risk and Expected Shortfall Forecasts When Many Methods Are Available

Value-at-risk (VaR) and expected shortfall (ES) have become widely used risk measures for daily portfolio returns. As a result, many methods now exist for forecasting the VaR and ES. These include GARCH-based modelling, approaches involving autoregressive quantile models, and methods incorporating measures of realised volatility. When multiple forecasting methods are available, an alternative to method selection is forecast combination. In this paper, we consider the

combination of VaR and ES forecasts when a large pool of forecasts is available. As there have been few studies in this area, we implement a variety of new combining methods. In terms of simplistic methods, in addition to the mean, the availability of many forecasts leads us to use the median and mode. As a complement to the previously proposed performance-based weighted combinations, we use regularisation to limit the risk of overfitting due to the large number of weights. By viewing the forecasts of VaR and ES from each method as the bounds of an interval forecast, we are able to apply interval forecast combining methods. These include forms of trimmed mean, and a method involving a mixture of the probability distributions inferred from the VaR and ES forecasts. Our empirical analysis used six stock indices and pools of different sizes. With 90 methods, we obtained particularly strong results for a trimmed mean approach, the mixtures method, and performance-based weighted combinations. However, greater accuracy resulted for a pool of just six methods, chosen to ensure diversity, with the best results produced by performance-based weighting.

Xiaoqian Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China

Online conformal inference for multi-step time series forecasting

We consider the problem of constructing distribution-free prediction intervals for multi-step time series forecasting, with a focus on the temporal dependencies inherent in multi-step forecast errors. We establish that the optimal h-step-ahead forecast errors exhibit serial correlation up to lag (h−1) under a general non-stationary autoregressive data generating process. To leverage these properties, we propose the Autocorrelated Multi-step Conformal Prediction (AcMCP) method, which effectively incorporates autocorrelations in multi-step forecast errors, resulting in more statistically efficient prediction intervals. This method ensures theoretical long-run coverage guarantees for multi-step prediction intervals, though we note that increased forecasting horizons may exacerbate deviations from the target coverage, particularly in the context of limited sample sizes. Additionally, we extend several easy-to-implement conformal prediction methods, originally designed for single-step forecasting, to accommodate multi-step scenarios. Through empirical evaluations, including simulations and applications to data, we demonstrate that AcMCP achieves coverage that closely aligns with the target within local windows, while providing adaptive prediction intervals that effectively respond to varying conditions.

Shanika Wickramasuriya, Monash University, Australia

Hierarchical forecasting: The role of information

In hierarchical forecasting, the process of forecast reconciliation transforms a set of "raw" forecasts, which do not satisfy the hierarchical aggregation constraints in the real data, into a set of "coherent" forecasts, which do satisfy those constraints. In this paper, we demonstrate that reconciliation algorithms adjust raw forecasts by adding a proportion of their "incoherency" to them, and naturally, if the raw forecasts are already coherent, then these algorithms do not modify the raw forecasts at all. However, if each forecast is based on a distinct information set, and we have access to historical data to estimate the accuracy of raw forecasts, there is scope for improving raw forecasts by combining the information that each one carries, even when they are already coherent. We propose a new method, called "information combination" (IComb) method that combines the information content of forecasts of different components in the process of reconciliation. We provide simulation evidence to illustrate the role of information sets in forecasting. We apply this algorithm to datasets that have been used in the literature and discuss the results.

IP69 — Recent Developments on AI and Statistics

June 13 (Saturday) 13:30–15:10 **Room:** 215 Program

Xuan Bi, University of Minnesota, United States

Distribution-Preserving Watermarking for Mixed-Typed Data

With the development of generative AI, watermark techniques are widely used in detecting the authenticity of AI-generated data and protecting the rights of users and creators. While it is already well applied in data types including imaging and text data, watermarking tabular data are still under-explored. Existing methods mostly focus on numerical data, leaving

discrete, categorical data, and mixed data less studied. In this work, we propose a novel framework for watermarking tabular data, which can accommodate numerical, discrete, categorical, and mixed data while keeping their distributions invariant. Furthermore, we also developed a corresponding detection mechanism, which can effectively detect watermark in the given data types above. We evaluate our method on multiple simulated and real datasets and demonstrate its effectiveness and robustness against shuffling and subsetting, while maintaining data utility.

Helen Zhang, University of Arizona, United States

Dynamic Supervised Principal Component Analysis for Classification

High-dimensional data classification is challenged by distributions that shift dynamically over time, making static subspace definitions and decision boundaries inadequate. We propose a novel framework for dynamic classification in high-dimensional spaces, designed to accommodate evolving class distributions across time or other index variables. The framework employs a supervised dimension reduction technique based on kernel smoothing to identify an optimal subspace and construct adaptive classification boundaries that respond to distributional changes. We develop theory and computational algorithms for both linear and quadratic discriminant analysis, and illustrate effectiveness of the proposed approach through simulation studies and real data applications.

Jian Shi, Chinese Academy of Sciences, China

TBD

TBD

Lei Li, Chinese Academy of Sciences, China

Empirical Lossless Compression Bound of a Data Sequence

We consider the lossless compression bound of an arbitrary individual data sequence. Conceptually, its Kolmogorov complexity provides such a bound, although it is uncomputable. According to Shannon's source coding theorem, the average compression bound is nH, where n is the number of words and H is the entropy of an oracle probability distribution characterizing the data source. The entropy obtained by plugging in the maximum likelihood estimate underestimates this bound. Shtarkov showed that the normalized maximum likelihood (NML) distribution is optimal in a minimax sense for any parametric family. By fitting a data sequence—without any a priori distributional assumption—using a relevant exponential family, we apply local asymptotic normality to derive the NML code length, which depends on Shannon entropy, dictionary size, and Fisher information. We further demonstrate that sequentially predicting the optimal code length for the next word via a Bayesian mechanism leads to a mixture code that also achieves the bound. The asymptotics apply not only to discrete symbols but also to continuous data, if the code length for the former is replaced by the description length for the latter. The analytical results are illustrated by calculating compression bounds for protein-encoding DNA sequences under different parsing models. Typically, compression is maximized when parsing aligns with amino acid codons, whereas pseudo-random sequences remain incompressible, as predicted by Kolmogorov complexity. Notably, the empirical bound becomes more accurate as the dictionary size increases.

DL03 — Semiparametric Modeling and Predictive Inference

June 13 (Saturday) 15:30–17:10 **Room:** LT1A Program

Huixia Judy Wang, Rice University, United States

Semiparametric Distribution Learning and Predictive Inference: A Quantile Regression Process Perspective

Modern data analysis increasingly requires learning not only average trends, but also heterogeneity, uncertainty, tail behavior, and how information can be fused across heterogeneous data sources. In this talk, I will discuss how the quantile regression process provides a flexible semiparametric approach to these problems by learning conditional distributions without imposing strong parametric assumptions on their shape.

I will highlight its role in several modern statistical problems, including multiple imputation, Bayesian inference, extreme quantile analysis, and conformal prediction, where quantile processes can help construct density-based nonconformity scores and prediction regions under complex error distributions. I will also discuss rank-based data integration motivated by the fusion of multiple epigenetic clocks for assessing biological aging. Together, these examples illustrate how quantile-based thinking can move beyond mean-centered modeling toward a richer and more robust understanding of variation, uncertainty, and individualized prediction.

Tony Sit, The Chinese University of Hong Kong, Hong Kong

Censored Quantile Regression with Time-Dependent Covariates and its Extensions

Quantile regression provides a robust framework for survival analysis, yet incorporating time-dependent covariates remains a significant challenge. This talk presents a unified framework for censored quantile regression with dynamic predictors. First, we introduce the base model for right-censored failure time data. This formulation generalises the definition of quantiles and provides a dynamic perspective for interpretation, extending the scope beyond independent censoring or longitudinal settings. Second, we extend this framework to recurrent events via a generalised quantile recurrence time model and competing risks by defining conditional quantiles through the cumulative incidence function, effectively managing dependent censoring and avoiding the non-identifiability issues inherent in marginal distributions. By integrating these three components, the talk provides a comprehensive toolkit for analysing modern time-to-event data where covariate effects evolve over time.

Seyoung Park, Yonsei University, South Korea

Smoothed Quantile Additive Regression with Functional Lasso Kernel Regularization in High Dimensions

We introduce kernel-based estimators tailored for high-dimensional additive quantile regression within the smooth backfitting (SBF) paradigm. To address the non-differentiability of the quantile check loss, we incorporate a smoothed quantile criterion and integrate it with a functional Lasso penalty, yielding a new sparsity-inducing regularization scheme. For computation, we design an alternating direction method of multipliers (ADMM) SBF procedure that decomposes the optimization into tractable subproblems, enabling parallel updates of additive components and enforcing roughness constraints exactly, which markedly enhances efficiency over classical smooth backfitting, especially in high dimensions. Our framework scales to ultra-high dimensional settings by combining smoothing and penalization, leading to consistent support recovery. We derive non-asymptotic risk bounds and establish variable-selection consistency under a projection-level incoherence condition. In addition, we construct a bias-corrected estimator, allowing principled statistical inference on individual component functions. The practical advantages of the method are demonstrated through comprehensive simulations and an application to genomic data, showing clear gains in both prediction and support identification compared to existing approaches for high-dimensional additive quantile regression.

DL10 — Advances in Nonparametric Statistics and Functional Data Analysis

June 13 (Saturday) 15:30–17:10 **Room:** LT1B Program

Aurore Delaigle, The University of Melbourne, Australia

Nonparametric estimators of nonstationary densities of streaming data

We consider nonparametric estimation of the nonstationary density of streaming data collected continuously over time. Those data are typically not entirely accessible at all times, and analyzing them requires dynamic approaches that do not require repeated access to past data. Several nonparametric estimators of nonstationary densities have been suggested in the literature, which all require choosing important tuning parameters at each time. We study theoretical properties of those estimators and propose a data-driven dynamic selection of their tuning parameters, which can be implemented iteratively and requires only sequential access to consecutive blocks of the most recent data, and which includes a selection of the sizes of the blocks. We illustrate the procedure through simulated and real streaming data.

Frédéric Ferraty, Toulouse Mathematics Institute, France

Nonparametric Approach to Forecasting Density-Valued Time Series

TBD

Dominik Liebl, University of Bonn, Germany

Making Event Study Plots Honest: A Functional Data Approach to Causal Inference

Event study plots are the centerpiece of Difference-in-Differences (DiD) analysis, but current plotting methods cannot provide honest causal inference when the parallel trends and/or no-anticipation assumptions fail. We introduce a novel functional data approach to DiD that directly enables honest causal inference via event study plots. Our DiD estimator converges to a Gaussian process in the Banach space of continuous functions, enabling powerful simultaneous confidence bands. This theoretical contribution allows us to turn an event study plot into a rigorous honest causal inference tool through equivalence and relevance testing: Honest reference bands can be validated using equivalence testing in the pre-treatment period, and honest causal effects can be tested using relevance testing in the post-treatment period. We demonstrate the performance of our method in simulations and two case studies.

IP16 — Recent Advances in Network Analysis and Dependence Testing for Complex Data

June 13 (Saturday) 15:30–17:10 **Room:** 209A Program

Fan Wang, The University of Melbourne, Australia

Smooth and abrupt changes in autoregressive tensor models.

In this talk, we propose an autoregressive tensor model, with time-dependent regression coefficients. With respect to the regression coefficients, three different regimes are considered: stationary, smooth-varying and abruptly-changing. We propose a computationally-efficient estimation procedure to handle these three regimes simultaneously, supported with theoretical guarantees and numerical experiments. Two extensions are considered: dynamic community detection in these three regimes and an estimation procedure for a more general class of time series models.

Zihan Wang, Tsinghua University, China

Time Series Gaussian Chain Graph Models

Time series graphical models have gained significant attention recently for representing complex dynamic dependence structure in multivariate time series. In this paper, we propose a new class of time series Gaussian chain graphs to capture blockwise temporal and cross-sectional dependence among component series that arise in various applications. Building on the AMP Markov property, the proposed model represents contemporaneous and lagged causal relations via directed edges across blocks, and the within-block conditional dependence via undirected edges. This formulation induces a group sparse plus group low-rank decomposition of the inverse spectral density across frequencies, which we exploit to establish novel identifiability conditions for recovering the time series chain graph. We further develop a frequency-domain learning procedure based on a regularized Whittle likelihood with two penalties: a group Lasso penalty to enforce group sparsity and a tensor-unfolding nuclear-norm penalty to capture a shared group low-rank structure. Under mild regularity conditions, we demonstrate that the proposed method achieves asymptotic consistency and exact edge recovery. The empirical performance is supported by extensive simulation studies and further demonstrated by an application to U.S. macroeconomic data that highlights key monetary-policy transmission mechanisms.

Qing Jiang, Beijing Normal University, China

Statistical inference for high-dimensional logistic regression with latent random change point

We study high-dimensional logistic regression with a latent random change point, where regression effects vary when a threshold covariate crosses an unobserved subject-specific change point governed by a low-dimensional parameter. This induces a mixture structure with a sparse high-dimensional regression component and a low-dimensional gating component, posing challenges for estimation, initialization, and hypothesis testing. We propose a unified framework to address these issues. We establish identifiability of the threshold-active submodel, develop a constructive initializer, and introduce a new high-dimensional EM truncation algorithm. Finally, we study the problem of testing for the presence of a change-point effect. Since the change-point distribution parameter is not identifiable under the null, standard methods fail. We construct a supremum score test and approximate its null distribution via multiplier bootstrap.

Qin Fang, The University of Sydney, Australia

Large-Scale Multiple Testing of Cross-Covariance Functions with Applications to Functional Network Models.

The estimation of functional networks through functional covariance and graphical models have recently attracted increasing attention in settings with high dimensional functional data, where the number of functional variables p is comparable to, and maybe larger than, the number of subjects. However, the existing methods all depend on regularization techniques, which make it unclear how the involved tuning parameters are related to the number of false edges. In this paper, we first reframe the functional covariance model estimation as a tuning-free problem of simultaneously testing p(p-1)/2 hypotheses for cross-covariance functions, and introduce a novel multiple testing procedure. We then explore the multiple testing procedure under a general error-contamination framework and establish that our procedure can control false discoveries asymptotically. Additionally, we demonstrate that our proposed methods for two concrete examples: the functional covariance model for discretely observed functional data and, importantly, the more challenging functional graphical model, can be seamlessly integrated into the general error-contamination framework, and, with verifiable conditions, achieve theoretical guarantees on effective false discovery control. Finally, we showcase the superiority of our proposals through extensive simulations and brain connectivity analysis of two neuroimaging datasets.

IP27 — Extremes, Heavy Tails and Stable Processes

June 13 (Saturday) 15:30–17:10 **Room:** 209B Program

Zaoli Chen, University of Science and Technology of China, China

Extremal Phase Transitions under Long-Range Dependence

We consider a family of stationary processes whose (1) marginal tail is sub-exponential (2) positive correlation satisfies a power-law decay without mean. By different combinations of marginal tails and correlation strengths, we can obtain drastically different extremal behaviors. In this talk, we shall focus on a type sub-exponentiality that lies in the Gumbel maximum domain of attraction with a moderate level of heaviness, e.g. the log-normal type of distributions. The goal is to illustrate delicate extremal phase transitions as the serial dependence gets stronger. The talk is based on a joint work with G. Samorodnitsky.

Bikramjit Das, Singapore University of Technology and Design, Singapore

Measuring extreme tail association

Simultaneous occurrences of extreme events need not imply symmetric or reciprocal tail dependence. However, most existing measures of extremal dependence are inherently symmetric and hence often fail to capture directional influence in tail association. We introduce a rank-based measure of Extreme Tail Association (ETA) for bivariate data quantifying such directional influence of one variable on another in extreme tail regions. The proposed estimator is easily computable, consistent with its population counterpart, and asymptotically normal under mild conditions, allowing for statistical inference. We further develop a formal test for asymmetry in tail association based on a multiplier bootstrap procedure. The practical relevance of the methodology is illustrated using data on extreme price movements in major cryptocurrencies. Beyond

providing a flexible tool for extremal association, the proposed framework offers a substantive argument for investigating causal relationships in extreme scenarios.

Ayan Bhattacharya, Indian Institute of Technology Bombay, India

Asymptotic behavior of extremes of regularly varying branching random walk on time-inhomogeneous random tree

In this talk, we study a branching random walk (BRW) on the real line whose genealogical structure is given by a supercritical branching process in an i.i.d. random environment satisfying the Kesten–Stigum condition. The displacements of offspring from a common parent are assumed to have jointly regularly varying tails. We investigate the quenched growth of the extremes and establish the weak convergence of the associated extremal processes formed by appropriately scaled particle positions in the n-th generation. We show that the limiting point process belongs to the class of randomly scaled, scale-decorated Poisson point processes (SScDPPP). As a consequence, an analogue of the predictions of Brunet and Derrida (2011) holds in this setting. Finally, we present a counterexample demonstrating that this framework—namely, weak convergence of the maximum position and of the extremal processes—cannot be extended to study the growth of extremes for branching random walks indexed by a Galton–Watson process in a deterministic environment. This is joint work with Zbigniew Palmowski.

IP30 — Statistical Methods for High-dimensional and Complex Data Structures

June 13 (Saturday) 15:30–17:10 **Room:** 203 Program

Kento Egashira, Tokyo University of Science, Japan

Change-point detection for mean and covariance structures in high-dimensional data under a strongly spiked eigenvalue model

We study change-point detection in high-dimensional data with low sample sizes, especially under a strongly spiked eigenvalue model. We propose a multivariate CUSUM-type statistic that simultaneously captures changes in mean vectors and covariance structures across potential change points. Under mild regularity conditions, we establish the consistency of the change-point estimator. In addition, we develop a change detection test based on the proposed statistic and derive its asymptotic behavior under the null hypothesis of no structural change, which enables asymptotically valid inference. Simulation studies demonstrate the robustness and favorable low-sample performance of the proposed approach in high-dimensional settings. We further illustrate its practical usefulness through an analysis of real data, confirming its effectiveness in detecting structural changes in complex multivariate data.

Anne van Delft, Columbia University, United States

Analyzing shape in a time series of random geometric objects

We introduce a new framework to analyze shape descriptors that capture the geometric features of an ensemble of point clouds. At the core of our approach is the point of view that the data arises as sampled recordings from a metric space-valued stochastic process, possibly of nonstationary nature, thereby integrating geometric data analysis into the realm of functional time series analysis. Our framework allows for natural incorporation of spatial-temporal dynamics, heterogeneous sampling, and the study of convergence rates. Further, we derive complete invariants for classes of metric space-valued stochastic processes in the spirit of Gromov, and relate these invariants to so-called ball volume processes. Under mild dependence conditions, a weak invariance principle in $D([0,1]\times[0,\mathscr{R}])$ is established for sequential empirical versions of the latter, assuming the probabilistic structure possibly changes over time. Finally, we use this result to introduce novel test statistics for topological change, which are distribution-free in the limit under the hypothesis of stationarity. We explore these test statistics on time series of single-cell mRNA expression data, using shape descriptors coming from topological data analysis.

Han Lin Shang, Macquarie University, Australia

Conformal prediction for high-dimensional functional time series

In statistics, forecast uncertainty is often quantified using a specified statistical model, though such approaches may be vulnerable to model misspecification, selection bias, and limited finite-sample validity. While bootstrapping can potentially mitigate some of these concerns, it is often computationally demanding. Instead, we take a model-agnostic and distribution-free approach, namely conformal prediction, to construct prediction intervals in high-dimensional functional time series. Among a rich family of conformal prediction methods, we study split and sequential conformal prediction. In split conformal prediction, the data are divided into training, validation, and test sets, where the validation set is used to select optimal tuning parameters by calibrating empirical coverage probabilities to match nominal levels; after this, prediction intervals are constructed for the test set, and their accuracy is evaluated. In contrast, sequential conformal prediction removes the need for a validation set by updating predictive quantiles sequentially via an autoregressive process. Using subnational age-specific log-mortality data from Japan and Canada, we compare the finite-sample forecast performance of these two conformal methods using empirical coverage probability and the mean interval score.

Anirvan Chakraborty, Indian Institute of Science Education and Research, Kolkata, India

Near-perfect Clustering Based on Recursive Binary Splitting Using Max-MMD

We develop novel clustering algorithms for functional data when the number of clusters K is unspecified and also when it is specified. These algorithms are developed based on the Maximum Mean Discrepancy (MMD) measure between the empirical distributions associated with two sets of observations. The algorithms recursively use a binary splitting strategy to partition the dataset into two subgroups such that they are maximally separated in terms of an appropriate weighted MMD measure. When K is unspecified, the proposed clustering algorithm has an additional step to check whether a group of observations obtained by the binary splitting technique consists of observations from a single population. We also learn K directly from the data using this algorithm. When K is specified, a modification of the previous algorithm is proposed which consists of an additional step of merging subgroups which are similar in terms of the weighted MMD distance. The theoretical properties of the proposed algorithms are investigated in an oracle scenario that requires the knowledge of the empirical distributions of the observations from different populations involved. In this setting, We prove that the algorithm proposed when K is unspecified achieves perfect clustering while the algorithm proposed when K is specified has the perfect order preserving (POP) property. Extensive real and simulated data analyses using a variety of models having location difference as well as scale difference show near-perfect clustering performance of both the algorithms which improve upon the state-of-the-art clustering methods for functional data.

IP26 — Advances in High-dimensional Statistical Inference and Dynamic Modeling

June 13 (Saturday) 15:30–17:10 **Room:** 201 Program

Shu-Chin Lin, National Taiwan University, Taiwan

Domain Selection for Functional Linear Models

This work addresses the domain selection problem in scalar-on-function linear regression. Although the functional predictor X(t) is observed over a compact domain, the scalar response Y is often associated with X(t) only within an unknown subregion. While Hall and Hooker (2016) proposed methods to estimate this region of association, accurately identifying the domain boundary remains challenging, particularly when the coefficient function tapers smoothly to zero. To address this limitation, we develop a reproducing kernel Hilbert space (RKHS) approach that specifically targets the domain selection problem. We introduce a novel domain estimator and establish its asymptotic properties under mild smoothness assumptions, demonstrating improved boundary identification even when the signal strength decays gradually.

Weichen Wang, The University of Hong Kong, Hong Kong

On reference-regulated multiperiod mean-variance portfolio optimization in high dimensions

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 estimation 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.

Lijia Wang, City University of Hong Kong, Hong Kong

Network autoregression for binary responses in social networks

Studying the propagation of binary responses on nodes in a large-scale social network is critical for understanding how individual behaviors and decisions are shaped by social structures and for predicting collective outcomes. We propose a network autoregressive model for binary-valued responses, in which the probability of response at each node is influenced by its neighbors' past decisions, its own past decision, and node-specific covariates, through a logistic link function. The model accounts for network noise and community structure by assuming the underlying network is generated from a block model, with autoregressive parameters that are community-specific. We establish conditions under which the long term behavior of the high-dimensional binary vector converges to a community-specific distribution and the associated convergence rate, illustrating when individuals in the same community or across the whole network reach a consensus regardless of their initial positions. Given an observed network and response vectors, we show asymptotic consistency and normality of the maximum likelihood estimators. We demonstrate the efficiency and validity of the inference procedure through simulated and real data. In particular, we show the model can be used to study the dynamics of strike occurrences in China and highlight the impact of online social network in facilitating collective actions.

Le Zhou, Hong Kong Baptist University, Hong Kong

Adaptive Testing and Estimation for High-Dimensional Covariance Change-Points

We build a series of tests based on U-statistics for testing the high-dimensional covariance matrix change-points. The asymptotic distributions of the constructed U-statistics are derived under the null and local alternative hypotheses. Then, we propose a family of maximum-type statistics, after which two test methods based on the combination of the p-values of these maximum-type statistics are developed. We also propose some methods to estimate the location of the change-point and obtain their corresponding convergence rates. In addition, three new adaptive estimations are built. Finally, the binary segmentation method is proposed to be combined with our three adaptive estimators to detect multiple change-points. Our simulation study shows that the proposed test methods can maintain high powers under alternatives with different sparsity levels and that our proposed adaptive estimators perform well under different alternatives with both single and multiple change-points.

IP33 — Recent Advances in Copula Theory and Applications

June 13 (Saturday) 15:30–17:10 **Room:** 202 Program

Alexander McNeil, York University, United Kingdom

Measures and Models of Non-Monotonic Dependence

We propose a margin-free measure of bivariate association generalizing Spearman's rho to the case of non-monotonic dependence that is defined in terms of two square integrable functions on the unit interval. We investigate properties of generalized Spearman correlation when the functions are piecewise continuous and strictly monotonic, with particular focus on

the special cases where the functions are drawn from orthonormal bases defined by Legendre polynomials and cosine functions. For continuous random variables, generalized Spearman correlation is treated as a copula-based measure and shown to depend on a pair of uniform-distribution-preserving (udp) transformations determined by the underlying functions. We derive bounds for generalized Spearman correlation and we use a novel technique that we refer to as stochastic inversion of udp transformations to construct singular copulas that attain the bounds and parametric copulas with densities that interpolate between the bounds and model different degrees of non-monotonic dependence. We also propose sample analogues of generalized Spearman correlation and investigate their asymptotic and small-sample properties. Potential applications of the theory are demonstrated including: exploratory analyses of the dependence structures of datasets and their symmetries; elicitation of functions maximizing generalized Spearman correlation via expansions in orthonormal basis functions; and construction of tractable probability densities to model a wide variety of non-monotonic dependencies.

Issey Sukeda, The University of Tokyo, Japan

Dependence modeling for circular data in neuroscience

Understanding statistical dependencies in neural oscillations, such as EEG phase, requires models that can deal with their inherently circular nature. In this work, we introduce two complementary approaches for dependence modeling of circular data in neural data. First, we introduce a torus graph modeling framework for multivariate phase data. By representing oscillatory phases across electrodes as points on a high-dimensional torus, this method estimates conditional dependence structures through a score matching. This approach enables identification of phase-specific functional connectivity while avoiding linearization of circular variables. We demonstrate that torus graphs capture frequency-specific coupling patterns and provide interpretable network structures. Second, we propose a unified statistical procedure for phase–amplitude coupling (PAC) that extends existing regression-based approaches to a joint modeling framework. Whereas conventional PAC metrics typically quantify dependence only in a pairwise (two-dimensional) fashion and cannot fully capture higher-order interactions, our circular–linear dependence model provides a likelihood-based approach that naturally generalizes to multivariate settings. This framework enables principled post-hoc inference and allows for modeling PAC strength in the context of higher-order (beyond bivariate) dependence structures.

Jia-Han Shih, National Sun Yat-sen University, Taiwan

A class of regression association measures based on concordance

Regression association measures aiming at predictability of a dependent variable Y from an independent variable X have received considerable attention recently. In this talk, we provide a unified discussion of some existing measures, including their rationale, properties, and estimation. Motivated by these measures, we consider a class of regression association measures which views the regression association of Y from X as the association of two independent replications from the conditional distribution of Y given X. We illustrate that the so-called Markov product copulas can be employed as a neat and convenient building block for this class of measures, and the measures so constructed can be expressed as a common form of the proportion of the variance of some function of Y that can be explained by X, rendering the measures a direct interpretation in terms of predictability. Also, the notion of two independent replications from the conditional distribution leads to a simple and elegant nonparametric estimation method based on the induced order statistics, hence, no smoothing techniques are required. Under the considered general framework, the performances and utilities of the regression association measures are examined through simulations and real data applications.

Martial Longla, University of Mississippi, United States

On estimation problems based on new types of copulas

New copula families are constructed with predetermined dependence structures . These copulas are studied in two setups. We consider a simple random sample from a bivariate distribution based on the copula family and a stationary Markov chain generated by these copulas. Copula parameter estimators are proposed along with their asymptotic distributions. Convex combinations are used to extend these copula families to copulas that exhibits tail dependence. Copula parameter estimation is considered for this extension and asymptotic distributions are provided. The Spearman's correlation coefficient and Kendall's coefficient of association are provided for each of these copula families. A two step estimation procedure is proposed.

IP37 — Additive Models and High-dimensional Inference: Modern Tools for Complex Data

June 13 (Saturday) 15:30–17:10 **Room:** 214 Program

Jeong Min Jeon, Seoul National University, South Korea

Hilbertian additive regression with general estimated variables.

In this talk, we introduce a Hilbertian additive model in which the response variable is Hilbert-space-valued and predictors are multi-dimensional Euclidean. We allow for the scenario where both variables are unobservable but they are estimable. This scenario includes various principal or singular component scores and density-valued responses. For such cases, we provide estimation errors for the variables, which are of importance in their own right. Additionally, we derive the full asymptotic properties of our regression estimator under such estimation errors. We demonstrate the strong performance of our regression estimator via simulation studies and a real data application.

Eun Ryung Lee, Sungkyunkwan University, South Korea

Parallel Additive Regression for High-Dimensional Data: Efficient Computation with Convergence and Consistency Guarantees

Smooth backfitting is a widely used estimation technique for additive regression models, but existing approaches face challenges in high-dimensional settings. Traditional iterative algorithms are computationally expensive due to sequential updates of component functions and the use of ℓ_1-based penalties often introduce significant estimation bias. To address these issues, we propose a novel nonconvex smooth backfitting algorithm for ultra-high dimensional additive models. Our method incorporates a two-step procedure: an initial parallel Lasso optimization for computational efficiency, followed by a weighted Lasso step guided by the SCAD penalty to reduce bias. This gradient-based algorithm allows parallel updates of component functions, significantly improving computational speed while maintaining theoretical guarantees. We establish that the proposed method achieves oracle estimation consistency and prove that the SCAD-based step identifies zero and non-zero components with perfect accuracy in a few iterations. Simulation studies demonstrate the superior finite-sample performance and scalability of our approach compared to existing methods.

Seong J. Yang, Jeonbuk National University, South Korea

A regularization approach for time-dependent AUC in survival analysis using deep neural networks

This study proposes a methodology for achieving robust discrimination performance in survival analysis. Typically, survival analysis is not sufficiently cautious of time-dependent performance. For example, the representative metric of performance, the C-index, does not account for fluctuations of performance over time, which can disrupt robust decision-making over time. In this study, we consider time-dependent analysis to assess the ability of a marker and propose an algorithm to mitigate fluctuations in predictive performance, as measured by AUC(t). With the help of a deep neural network, its optimizers, and surrogate penalty functions, we achieve a more robust performance over time. We examine the model's discriminative power using AUC(t) and the C-index on synthetic datasets and three real datasets. Results show that the proposed algorithm can provide more robust and improved discrimination over time, which is useful when performance variation affects the decision-making along with the timeline.

Ming-Yen Cheng, Hong Kong Baptist University, Hong Kong

Sparse optimal model averaging under a general framework

We introduce a general high-dimensional model averaging framework that applies to various settings. The commonly imposed but restrictive assumptions of (i) correctly specified full model, (ii) homoscedasticity/specific variance function and (iii) finite/low number of candidate models are simultaneously relaxed. Conventional model averaging methods use a LASSO-type penalty on model complexity; however, this does not guarantee sparsity of selected weights in finite samples. We tackle this crucial and challenging problem under the suggested general framework. Firstly, we study testing

significance of a group of candidate models and suggest a novel test statistic. Notably, it enjoys the Wilks phenomenon and immediately leads to an asymptotically distribution-free test that can detect root-n local alternatives. Then, we employ the proposed test as a building block to develop a data-driven sparse optimal model weight selection method. We show that it possesses the desirable weight sparsity, in that it can recover the sparse structure of the infeasible population optimal weights, and asymptotic optimality properties. Extensive simulations demonstrate the robustness and superior performance of our methodology over existing alternatives. A data analysis illustrates its utility in applications.

DL18 — Advancing Precision Health with AI and Statistics

June 14 (Sunday) 10:20–12:00 **Room:** LT1A Program

Hsin-Chou Yang, Academia Sinica, Taiwan

Unlocking Precision Health: Insights from Genetics, Medical Imaging, and Multimodal AI Integration

TBD

Jung-Ying Tzeng, North Carolina State University, United States

Statistical learning for polygenic risk prediction and CNV association testing

Modern genomic analyses face growing challenges due to population diversity and restrictions on sharing individual-level data. This talk presents two statistical learning approaches that address these challenges by leveraging information across studies while accounting for heterogeneity and privacy constraints. (1) The first part focuses on a transfer learning approach for polygenic risk score construction when base data and target data may differ in ancestral background. By treating genome-wide association study (GWAS) summary statistics from the base data as knowledge learned from a pre-trained model, we adopt a transfer learning framework to effectively leverage the knowledge learned from the base data to build prediction models for target individuals. Our proposed framework consists of two main steps: (i) conducting false-negative-control marginal screening to extract useful knowledge from the base data; and (ii) performing joint model training to integrate the extracted knowledge with target training data for trans-data prediction. The proposed approach substantially enhances both computational and statistical efficiency of joint-model training, and yields more accurate trans-data prediction across a wide range of heterogeneity levels between target and base data. (2) The second part focuses on a federated learning approach for rare copy number variant (CNV) association testing. Rare CNV analysis requires large sample sizes, and approaches based on summary statistics, such as meta-analysis or federated learning, provide efficient and privacy-preserving strategies for data integration. However, unlike SNPs, CNV meta-analysis faces additional challenges that complicate cross-study integration, including inconsistent CNV locus definitions across studies and the lack of standardized association summary outputs. We adopt a federated learning framework for rare CNV association testing where a coordinating site collects study-level summary statistics and CNV genomic information. We develop a federated testing algorithm that harmonizes CNV locus definitions across studies, computes the collapsing test statistics and p-values based on the rare-CNV test CONCUR, and reduces communication to a single round. The method accommodates study-specific covariates with effects varying across studies, a scenario often encountered in multi-ethnic analyses. For both works, we demonstrate the utilities and scalability of the proposed approaches using simulations and real data applications.

Tso-Jung Yen, Academia Sinica, Taiwan

An Automatic Approach to Explainable AI with Applications to Medical Image Classification

Recent advances in explainable AI have inspired researchers to develop methods that can effectively detect input impact on the output of a predictive model. However, most of these methods require multiple stages to process the input. Some of the stages involve re-sampling the input data. When there are large amounts of input data, such a re-sampling procedure may take time to proceed. In this talk, we propose a method that can automatically detect input impact on the output of the predictive model. Our method relies on the idea that turns the impact detection problem to a prediction problem. Simulation experiments show that our method can correctly carry out the input impact detection without spending too much time on input processing.

Keywords: Deep learning; Explainable AI; Local Interpretable Model-agnostic Explanations; Ultrasound images.

DL11 — Statistical Inference in Metric Spaces

June 14 (Sunday) 10:20–12:00 **Room:** LT1B Program

Xueqin Wang, University of Science and Technology of China, China

Metric Distribution Function: A New Statistical Cornerstone for Non-Euclidean Data

This report introduces the Metric Distribution Function (MDF)—a breakthrough framework that transcends the limitations of Euclidean space. By unifying mainstream non-parametric methods and supporting statistical inference for data with non-Euclidean geometries, the MDF provides a new theoretical and practical foundation for the analysis of large-scale, complex-structured data.

Ting Li, Southern University of Science and Technology, China

Ball Impurity: Measuring Heterogeneity in General Metric Spaces

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

Jin Zhu, University of Birmingham, United Kingdom

Identification of Genetic Factors Associated with Corpus Callosum Morphology: Conditional Strong Independence Screening for Non-Euclidean Responses

The corpus callosum, the largest white matter structure in the brain, plays a critical role in interhemispheric communication. Variations in its morphology are associated with various neurological and psychological conditions, making it a key focus in neurogenetics. Age is known to influence the structure and morphology of the corpus callosum significantly, complicating the identification of specific genetic factors that contribute to its shape and size. We propose a conditional strong independence screening method to address these challenges for ultrahigh-dimensional predictors and non-Euclidean responses. Our approach incorporates prior knowledge, such as age. It introduces a novel concept of conditional metric dependence, quantifying non-linear conditional dependencies among random objects in metric spaces without relying on predefined models. We apply this framework to identify genetic factors associated with the morphology of the corpus callosum. Simulation results demonstrate the efficacy of this method across various non-Euclidean data types, highlighting its potential to drive genetic discovery in neuroscience.

IP40 — Advances in Statistical Machine Learning

June 14 (Sunday) 10:20–12:00 **Room:** 209A Program

Ben Dai, The Chinese University of Hong Kong, Hong Kong

EnsLoss: Stochastic Calibrated Loss Ensembles for Preventing Overfitting in Classification

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.

Hui Zou, University of Minnesota, United States

Double Descent in the Enhanced Response Envelope Model

The response envelope model provides substantial efficiency gains over the standard multivariate linear regression by identifying the material part of the response to the model and by excluding the immaterial part. In this paper, we propose the enhanced response envelope by incorporating a novel envelope regularization term based on a nonconvex manifold formulation. The enhanced response envelope naturally handles high-dimensional data for which the original response envelope is not serviceable without necessary remedies. In an asymptotic high-dimensional regime where the ratio of the number of predictors over the number of samples converges to a non-zero constant, we characterize the risk function and reveal an interesting double descent phenomenon for the envelope model. A simulation study confirms our main theoretical findings.

Yufeng Liu, University of Michigan, United States

Low-rank Reinforcement Learning with Heterogeneous Human Feedback

Modern decision-making systems, from online marketplaces to large language models (LLMs), increasingly rely on high-dimensional human feedback, where heterogeneous user preferences and massive feature spaces pose major challenges for statistical efficiency and alignment. In this talk, I will present low-rank reinforcement learning (RL) methods that exploit latent structures in human feedback to enable scalable and theoretically grounded learning. In the first part, we study the dynamic assortment problem in high-dimensional e-commerce and show how a low-rank structure in user–item interactions reduces the complexity of estimating personalized utilities and enables efficient exploration–exploitation strategies with provable regret guarantees. In the second part, we extend these ideas to reinforcement learning from human feedback (RLHF) in large-scale contextual environments, proposing a low-rank contextual framework that accommodates diverse user preferences and complex latent spaces in LLMs while providing theoretical guarantees on sample efficiency and robustness under distribution shifts.

Yoonkyung Lee, The Ohio State University, United States

Data Influence Dynamics under Iterative Training Algorithms

Influence functions are well-known tools for tracing model behavior back to the training data and provide a first order approximation of data influence measures. However, when models are trained with iterative algorithms in a challenging optimization landscape, parameter estimates may not reach exact optima. In these situations, the assumption of optimality and convexity in canonical influence analysis may lead to a notable discrepancy between the theoretical influence function and its numerical approximations. To overcome this limitation, we propose a new framework for influence analysis that explicitly accounts for the training trajectory and characterizes how data influence evolves during the training process. We apply the proposed framework to analyze the influence functions under popular training algorithms, including stochastic gradient descent (SGD), SGD with momentum, and proximal gradient descent. This reveals a recursive structure of data influence dynamics induced by iterative optimization. Numerical studies show that our approach consistently provides

more accurate estimates of data influence than the canonical approach in various settings, and that examining data influence dynamics enables more fine-grained model diagnostics.

IP03 — Advances in Graph Learning and Network Analysis

June 14 (Sunday) 10:20–12:00 **Room:** 209B Program

Jie Peng, University of California, Davis, United States

Inferring Latent Graphs from Stationary Signals

Graphs offer an intuitive way to represent relationships among variables. However, in many practical settings, the underlying graph is not directly observable and must be inferred from data. In this talk, we introduce a novel framework for inferring a latent graph under the assumption that the observed multidimensional signals are stationary with respect to the graph. We first formalize the notion of graph stationarity, and then propose a framework in which the covariance matrix of the observed signals is jointly diagonalizable with the normalized graph Laplacian matrix of the latent graph. We consider both the i.i.d. setting and the more general case of temporally correlated samples. Applications to S&P 500 stock price data and California temperature data demonstrate that the proposed framework yields efficient signal representations and leads to the construction of meaningful and interpretable graphs.

Wanjie Wang, National University of Singapore, Singapore

Node Differential Privacy in Node Ranking for Social Networks

Ranking nodes in a network is a fundamental task in data analysis, yet publishing such rankings can expose sensitive structural information about individual participants. We study node-differentially private (node-DP) algorithms for two canonical ranking problems: PageRank and the HITS authority score. Node-DP is a strong privacy guarantee that hides the entire set of outgoing connections of any single node, and is considerably harder to achieve than the more commonly studied edge-DP.

For PageRank, our approach is based on a graph projection technique: given a minimum-degree parameter m, we augment the input graph with m pseudo-nodes so that every node achieves out-degree exactly m, directly bounding the sensitivity of the PageRank vector under one-node perturbations. Adding calibrated Gaussian noise then yields an (ε, δ)-DP mechanism with per-component mean-squared error $O\big(\beta^2/[mn\varepsilon^2(1-\beta)^2]\big)$, at the cost of an approximation bias bounded by $O\big(\beta(1-d_{\min}/m)/(1-\beta)^2\big)$, which vanishes when m equals the minimum out-degree of the original graph. For the HITS authority score we employ the efficient Propose-Test-Release (ePTR) framework, which privately tests whether the spectral gap of $A^\top A$ is large enough to guarantee stable recovery. We exploit the Davis–Kahan perturbation theorem to bound the change in the leading right singular vector of the adjacency matrix, and show that a spectral gap condition on $A^\top A$ suffices for node-DP release with mean-squared error $O(1/[c^2 n\varepsilon^2])$.

Together, these results provide the first node-DP algorithms for network ranking with provable utility guarantees, closing a significant gap in the differential privacy literature on graph statistics.

Xin Tong, The University of Hong Kong

Stance Drift in AI-mediated communication

As large language models (LLMs) increasingly mediate communication—from drafting emails to summarizing scientific reports—a quiet risk emerges: the stance of a human message can change in transit. We study this AI-mediated scenario as a two-step generation–extraction process and introduce the stance preservation rate (SPR) to measure how well models retain original stances.

Hao Chen, University of California, Davis, United States

Community detection across mixing patterns for two or more communities

Community structure in networks can arise through a variety of mixing patterns, including assortative mixing, disassortative mixing, core-periphery structure, and combinations of these patterns across multiple communities. This talk presents a

unified framework for community detection across such settings, built on standardized edge-count statistics and a recursive bi-partitioning strategy. The two-community case serves as the basic building block, providing criteria for distinguishing different mixing structures and selecting an appropriate splitting rule. Building on this foundation, the framework extends to multi-community networks by recursively partitioning subnetworks, allowing different parts of the network to exhibit different forms of organization. Simulation studies and real-data examples illustrate the performance of the proposed approach and its ability to recover interpretable community structures in networks with heterogeneous mixing patterns.

IP10 — Statistics and AI

June 14 (Sunday) 10:20–12:00 **Room:** 203 Program

Ilsang Ohn, Inha University, South Korea

Adaptive online Bayesian inference via expert aggregation

In online settings where data arrive sequentially and the environment may shift, standard Bayesian inference can be more sensitive to the choice of learning rate and prior. We propose an online variational Bayes method that aggregates a pool of (generalized) Bayesian experts, each specified by a distinct learning rate or prior, using exponential-weights with a mixing correction. The aggregated predictor adaptively selects the best expert configuration without prior knowledge of the environment, and we establish cumulative regret bounds showing competitiveness with the best fixed expert up to a data-dependent tracking overhead. We further apply the framework to online conformal prediction, proving convergence of the interval width to the oracle quantile. Experiments on both synthetic and real data confirm that the method yields substantially more stable predictions than existing online variational and conformal baselines.

Masaaki Imaizumi, The University of Tokyo, Japan

High-dimensional theory for dynamics of neural network training and transformer inference

Toward theoretical understanding of deep learning and artificial intelligence, the dynamics of inference/traning of has been regarded as a key factor. The analysis for these dynamics leverages the recent development of physics-oriented high-dimensional theory for neural networks. While neural networks exhibit complex dynamics in many aspects, employing the high-dimensional limit to reduce it to the dynamics of element distributions enables effective analysis. The first topic describes several approaches for analyzing the training dynamics of deep neural networks, followed by an estimation of generalization error estimation and multi time-scales for feature unlearning. The second topic explains the research background of representing transformer inference using nonlinear dynamical models with coupled oscillators, and analyzes the mechanism by which this model induces degeneracy.

Kiseop Lee, Purdue University, United States

Attention based reading, highlighting, and forecasting of the limit order book

Managing high-frequency data in a limit order book (LOB) is complex due to its high dimensionality, irregular timing, and complex spatiotemporal dependencies across price levels. These challenges often exceed the capabilities of conventional time-series models. Accurate prediction of the multi-level LOB, not just the mid-price, is crucial for understanding market dynamics but is difficult due to the interdependencies among attributes like order types, features, and levels. This study introduces advanced sequence-to-sequence models to forecast the entire multi-level LOB, including prices and volumes. Our key contribution is a compound multivariate embedding method that captures spatiotemporal relationships. Empirical results show that our method outperforms others, achieving the lowest forecasting error while maintaining LOB structure.

IP17 — Statistical Learning Theory with Dependent/Markov Observations

June 14 (Sunday) 10:20–12:00 **Room:** 201 Program

Azadeh Khaleghi, ENSAE Paris, France

On the Estimation of Mixing Coefficients from Stationary Ergodic Sample Paths

Many problems in modern statistics and machine learning involve data with non-negligible temporal dependence, where classical i.i.d. assumptions break down and finite-sample guarantees depend explicitly on the strength of dependence. Mixing coefficients such as α-, β-, and ϕ-mixing provide a natural way to quantify this dependence and underpin much of the theoretical analysis for time series, stochastic processes, and online learning. However, these coefficients are typically treated as abstract assumptions and are rarely known or verifiable in practice. This motivates the study of their estimation from a single dependent sample-path. In this talk, I will discuss results on estimating α- and β-mixing coefficients in both continuous and finite-state settings, including nonparametric estimators with explicit convergence rates and strongly consistent estimators enabling goodness-of-fit testing for dependence strength. I will then briefly connect these ideas to restless bandit problems with dependent rewards, where finite-time regret guarantees rely critically on mixing assumptions, and conclude by highlighting open challenges that arise when dependence must be estimated or controlled in adaptive, data-driven settings.

Geoffrey Wolfer, Waseda University, Japan

Optimistic Estimation of Convergence in Markov Chains with the Average-Mixing Time

The convergence rate of a Markov chain to its stationary distribution is typically assessed using the concept of total variation mixing time. However, this worst-case measure often yields pessimistic estimates and is challenging to infer from observations. In this paper, we advocate for the use of the average-mixing time as a more optimistic and demonstrably easier-to-estimate alternative. We further illustrate its applicability across a range of settings, from two-point to countable spaces, and discuss some practical implications. Joint work with Pierre Alquier.

Vahe Karagulyan, ESSEC Business School, France

Empirical PAC-Bayes bounds for Markov chains

The core of generalization theory was developed for independent observations. Some PAC and PAC-Bayes bounds are available for data that exhibit a temporal dependence. However, there are constants in these bounds that depend on properties of the data-generating process: mixing coefficients, mixing time, spectral gap... Such constants are unknown in practice. In this paper, we prove a new PAC-Bayes bound for Markov chains. This bound depends on a quantity called the *pseudo-spectral gap*, γ_{ps}. The main novelty is that we can provide an empirical bound on γ_{ps} when the state space is finite. Thus, we obtain the first fully empirical PAC-Bayes bound for Markov chains. This extends beyond the finite case, although this requires additional assumptions. On simulated experiments, the empirical version of the bound is essentially as tight as the one that depends on γ_{ps}.

Daniel Paulin, Nanyang Technological University, Singapore

Scalable MCMC methods for Bayesian learning of time series models

Bayesian inference for time series models is often challenging due to the ill-conditioning of the posterior distribution. In this talk, we will introduce some new MCMC methods that perform well in such scenarios.

IP05 — Distribution Shift and Data Integration: State of the Art and Future Outlook

June 14 (Sunday) 10:20–12:00 **Room:** 202 Program

Jiwei Zhao, University of Wisconsin–Madison, United States

Towards the Efficient Inference by Incorporating Automated Computational Phenotypes under Covariate Shift

Collecting gold-standard phenotype data via manual extraction is typically labor-intensive and slow, whereas automated computational phenotypes (ACPs) offer a systematic and much faster alternative. However, simply replacing the gold-standard with ACPs, without acknowledging their differences, could lead to biased results and misleading conclusions. Motivated by the complexity of incorporating ACPs while maintaining the validity of downstream analyses, in this paper, we consider a semi-supervised learning setting that consists of both labeled data (with gold-standard) and unlabeled data (without gold-standard), under the covariate shift framework. We develop doubly robust and semiparametrically efficient estimators that leverage ACPs for general target parameters in the unlabeled and combined populations. In addition, we carefully analyze the efficiency gains achieved by incorporating ACPs, comparing scenarios with and without their inclusion. Notably, we identify that ACPs for the unlabeled data, instead of for the labeled data, drive the enhanced efficiency gains. To validate our theoretical findings, we conduct comprehensive synthetic experiments and apply our method to multiple real-world datasets, confirming the practical advantages of our approach.

Seong-ho Lee, University of Seoul, South Korea

Semiparametric Framework for Efficient Semi-supervised Inference under Label Shift

In many real-world applications, researchers aim to deploy models trained in a source domain to a target domain, where obtaining labeled data is often expensive, time-consuming, or even infeasible. While most existing literature assumes that the labeled source data and the unlabeled target data follow the same distribution, distribution shifts are common in practice. This paper focuses on label shift and develops efficient inference procedures for general parameters characterizing the unlabeled target population. A central idea is to model the outcome density ratio between the labeled and unlabeled data. To this end, we propose a progressive estimation strategy that unfolds in three stages: an initial heuristic guess, a consistent estimation, and ultimately, an efficient estimation. This self-evolving process is novel in the statistical literature and of independent interest. We also highlight the connection between our approach and prediction-powered inference (PPI), which uses machine learning models to improve statistical inference in related settings. We rigorously establish the asymptotic properties of the proposed estimators and demonstrate their superior performance compared to existing methods. Through simulation studies and multiple real-world applications, we illustrate both the theoretical contributions and practical benefits of our approach.

Molei Liu, Peking University, China

Transfer Learning of CATE with Kernel Ridge Regression

The proliferation of data has sparked significant interest in leveraging findings from one study to estimate treatment effects in a different target population without direct outcome observations. However, the transfer learning process is frequently hindered by substantial covariate shift and limited overlap between (i) the source and target populations, as well as (ii) the treatment and control groups within the source. We propose a novel method for overlap-adaptive transfer learning of conditional average treatment effect (CATE) using kernel ridge regression (KRR). Our approach involves partitioning the labeled source data into two subsets. The first one is used to train candidate CATE models based on regression adjustment and pseudo-outcomes. An optimal model is then selected using the second subset and unlabeled target data, employing another pseudo-outcome-based strategy. We provide a theoretical justification for our method through sharp non-asymptotic MSE bounds, highlighting its adaptivity to both weak overlaps and the complexity of CATE function. Extensive numerical studies confirm that our method achieves superior finite-sample efficiency and adaptability. We conclude by demonstrating the effectiveness and superiority of our approach on two real-world datasets.

Chi-Shian Dai, National Cheng Kung University, Taiwan

Multiclass Classification Utilizing Heterogeneous External Machine Learning Predictions

Classification is a fundamental task in medical research, where investigators seek not only accurate prediction but also interpretable relationships between risk factors and clinical outcomes. Multinomial logistic regression provides a principled inferential framework, with coefficients interpretable as log−relative risks. In contrast, modern machine-learning methods such as gradient boosting, XGBoost, Bayesian additive regression trees, and deep learning often achieve superior predictive accuracy but yield limited scientific interpretability. This paper proposes a unified framework that leverages summary-level nonparametric machine-learning predictions to strengthen an interpretable multinomial logistic regression model fitted at

the individual level in a primary study. We develop an empirical-likelihood-based approach that integrates these external predictions while explicitly addressing key challenges: external models may provide only partial information on outcome classes and covariates, and the primary and external studies may exhibit both proportion heterogeneity (differences in class prevalences) and feature heterogeneity (covariate shift). We establish consistency and asymptotic normality of the proposed estimator, characterize its efficiency, and investigate the impact of the quality of external predictions.

IP11 — Statistical and Algorithmic Foundation of Diffusion Models

June 14 (Sunday) 10:20–12:00 **Room:** 214 Program

Yuxin Chen, University of Pennsylvania, United States

Towards a Unified Framework for Guided Diffusion models

Guided or controlled data generation with diffusion models has become a cornerstone of modern generative modeling. Despite substantial advances in diffusion model theory, the theoretical understanding of guided diffusion samplers remains severely limited. We make progress by developing a unified algorithmic and theoretical framework that accommodates both diffusion guidance and reward-guided diffusion. Aimed at fine-tuning diffusion models to improve certain rewards, we propose injecting a reward guidance term – constructed from the difference between the original and reward-reweighted scores – into the backward diffusion process, and rigorously quantify the resulting reward improvement over the unguided counterpart. As a key application, our framework shows that classifier-free guidance (CFG) decreases the expected reciprocal of the classifier probability, providing the first theoretical characterization of the specific performance metric that CFG improves for general target distributions. When applied to reward-guided diffusion, our framework yields a new sampler that is easy-to-train and requires no full diffusion trajectories during training. Numerical experiments further corroborate our theoretical findings.

Andre Wibisono, Yale University, United States

Optimal Score Estimation via Empirical Bayes Smoothing

We study the problem of estimating the score function of an unknown probability distribution from independent and identically distributed observations in high dimensions. Assuming that the distribution is subgaussian and has a Lipschitz-continuous score function, we establish the optimal error rate for this estimation problem under the L_2 loss function that is commonly used in the score matching literature. Leveraging key insights in empirical Bayes theory as well as a new convergence rate of smoothed empirical distribution in Hellinger distance, we show that a regularized score estimator based on a Gaussian kernel attains this rate, shown optimal by a matching minimax lower bound. We discuss implication of our result on the sample complexity of score-based generative models.

Yuejie Chi, Yale University, United States

Polynomial Convergence of Riemannian Diffusion Models

Diffusion generative models have demonstrated remarkable empirical success in the recent years and are now considered the state-of-the-art generative models in modern AI. These models consist of a forward process, which gradually diffuses the data distribution to a noise distribution spanning the whole space, and a backward process, which inverts this transformation to recover the data distribution from noise. Most of the existing literature assumes that the underlying space is Euclidean. However, in many practical applications, the data are constrained to lie on a submanifold of Euclidean space. Addressing this setting, de Bortoli et al. (2022) introduced Riemannian diffusion models and proved that using an exponentially small step size yields small sampling error in Wasserstein distance, provided the data distribution is smooth and strictly positive. In this work, we prove that a polynomially small stepsize suffices to guarantee small sampling error in total variation distance, without any assumption on the smoothness or positivity of the data distribution. Our analysis only requires mild and standard curvature assumptions on the underlying manifold. Our approach opens the door to a sharper analysis of diffusion models on non-Euclidean spaces.

Arnak Dalalyan, ENSAE/CREST, France

Discretisation error of Denoising Diffusions measured in Wasserstein Distance

Generative modeling aims to produce new random examples from an unknown target distribution, given access to a finite collection of examples. Among the leading approaches, denoising diffusion probabilistic models (DDPMs) construct such examples by mapping a Brownian motion via a diffusion process driven by an estimated score function. In this work, we first provide empirical evidence that DDPMs are robust to constant-variance noise in the score evaluations. We then establish finite-sample guarantees in Wasserstein-2 distance that exhibit two key features: (i) they characterize and quantify the robustness of DDPMs to noisy score estimates, and (ii) they achieve faster convergence rates than previously known results. Furthermore, we observe that the obtained rates match those known in the Gaussian case, implying their optimality.

Joint with V. Arsenyan and E. Vardanyan https://arxiv.org/abs/2506.09681

IP23 — Advances in Causal Inference and Statistical Testing

June 14 (Sunday) 10:20–12:00 **Room:** 215 Program

Yifan Cui, Zhejiang University, China

Double Machine Learning of Continuous Treatment Effects with General Instrumental Variables

Estimating causal effects of continuous treatments is a common problem in practice, for example, in studying dose-response functions. Classical analyses typically assume that all confounders are fully observed, whereas in real-world applications, unmeasured confounding often persists. In this article, we propose a novel framework for local identification of dose-response functions using instrumental variables, thereby mitigating bias induced by unobserved confounders. We introduce the concept of a uniform regular weighting function and consider covering the treatment space with a finite collection of open sets. On each of these sets, such a weighting function exists, allowing us to identify the dose-response function locally within the corresponding region. For estimation, we develop an augmented inverse probability weighting score for continuous treatments under a debiased machine learning framework with instrumental variables. We further establish the asymptotic properties when the dose-response function is estimated via kernel regression or empirical risk minimization. Finally, we conduct both simulation and empirical studies to assess the finite-sample performance of the proposed methods.

Guoyu Zhang, Peking University, China

Bidirectional causal inference: a nonparametric potential outcome framework

Identifying causal effects is essential across various scientific disciplines. While much of the existing literature focuses on unidirectional causal effects, bidirectional relationships are common in real-world systems and often involve greater complexity. Previous research has mainly examined this scenario under parametric structural assumptions, but such models are often restrictive and difficult to generalize. This paper introduces a fully nonparametric potential outcome framework to analyze general bidirectional causal effects. Our approach accommodates both discrete and continuous variables and provides nonparametric identification results under mild assumptions. Based on these findings, we propose the corresponding nonparametric estimators. Extensive simulations confirm the strong finite-sample performance of the proposed method.

Xiaoyu Hu, Xian Jiaotong University, China

Neural Wasserstein Two-Sample Tests

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.

DL09 — To be confirmed

June 14 (Sunday) 13:00–14:40 **Room:** LT1A Program

Zhiliang Ying, Columbia University, United States

TBD

TBD

Yunxiao Chen, The London School of Economics and Political Science, United Kingdom

TBD

TBD

Jing Ouyang, The University of Hong Kong, Hong Kong

Statistical Analysis of Large-scale Item Response Data under Measurement Non-invariance: A Statistically Consistent Method and Its Application to PISA 2022

International Large-scale Assessments (ILSAs) collect valuable data on education quality and performance development across countries, enabling country groups to share effective techniques and policies. A key analytical tool for ILSAs is the Item Response Theory (IRT) model, which estimates performance distributions and group rankings. However, a major challenge in IRT calibration is that some items suffer from Differential Item Functioning (DIF), where different groups have different probabilities of correctly answering the items, controlling for individual proficiency. DIF is particularly common in ILSA due to the differences in test languages, cultural contexts, and curriculum designs across different groups. Ignoring or improperly accounting for DIF when calibrating the IRT model can lead to severe biases in the estimated performance distributions, which may further distort the ranking of the groups. Unfortunately, existing methods cannot guarantee the statistically consistent recovery of the group ranking without unrealistic assumptions for ILSA, such as the existence and knowledge of reference groups and anchor items. To fill this gap, this paper proposes a new approach to DIF analysis across multiple groups. This approach is computationally efficient and statistically consistent, without making strong assumptions about reference groups and anchor items. The proposed method is applied to PISA 2022 data from the mathematics, science, and reading domains, providing insights into their DIF structures and performance rankings of countries.

DL16 — Nonstandard Asymptotics in Causal Mediation Analysis

June 14 (Sunday) 13:00–14:40 **Room:** LT1B Program

Ian McKeague, Columbia University, United States

On accurately calibrating test statistics at singularities in multidimensional parameter spaces

Non-regular limit behavior of Wald-type test statistics at singularities in multidimensional parameter spaces is a common phenomenon. This typically leads to difficulties in accurately calibrating such tests due to instability of the test statistic in the neighborhood of the singularity. For example, the classic Sobel test for the presence of a causal mediation effect

involves testing whether the product of two regression parameters vanishes, giving rise to a N(0,1/4) limit in one part of the null and a N(0,1) limit in the other part. A systematic study of the limiting behavior of Wald-type tests involving such singularities was recently carried out by Dufour et al. (Annals of Statistics, 2025). The present talk introduces and discusses a way of calibrating test statistics in such settings to produce stable limit behavior at the singularity. An application to a post-selection inference problem arising in high-dimensional causal mediation analysis will be presented in detail.

Moulinath Banerjee, University of Michigan, United States

'HARMLESS' sampling for determining level sets of a response function

We propose a two-stage design for estimating a root or boundary defined by an unknown regression function observed with noise. In many applications, including dose-finding clinical trials, it is desirable to learn the zero level set while allocating observations mainly on one *safe* side of the unknown boundary. Classical one-sided stochastic approximation procedures guarantee finite overshoot but achieve convergence rates governed by the imposed bias, which are slower than the parametric rate.

In Stage I, a one-sided stochastic approximation is used to obtain a conservative preliminary estimator while preserving finite overshoot. In Stage II, additional observations are collected locally on the safe side and a local regression estimator is constructed. We show that the resulting estimator is asymptotically normal with the parametric rate, despite the finite-overshoot constraint. We further show that a variant of Polyak–Juditsky averaging fails to recover the parametric rate under finite-overshoot designs. Extensions to multivariate boundary estimation are developed for both linear and smooth nonparametric level sets. To demonstrate the effectiveness of our procedure for risk-controlled decision making, we conduct extensive simulations and apply it to a bank loan dataset to estimate the maximal borrower feature level that keeps the default probability below a prescribed threshold.

Guosheng Yin, The University of Hong Kong, Hong Kong

Bayesian Knockoff Filter for Controlling False Discovery Rate

In many scientific fields, researchers are interested in discovering features with substantial effects on the response from a large number of candidate features while controlling the proportion of false discoveries. By incorporating the knockoff procedure in a Bayesian framework, we develop the Bayesian knockoff filter (BKF) for selecting important features. BKF is built upon a tight upper bound for the posterior probability of the null hypothesis, which was derived under the Bayesian "flip-sign" property implied by our procedure. In contrast to the standard knockoff method in the frequentist paradigm, we allow the knockoff variables to be generated repeatedly in a Bayesian framework, which removes the unwanted variability of the single-knockoff generation (knockoff generated only once). Compared to multi-knockoff approaches, BKF is more principled in incorporating all relevant information. Numerical experiments on both synthetic and real data demonstrate the advantages of BKF over existing knockoff methods and Bayesian variable selection approaches, i.e., the BKF possesses higher power and yields a lower false discovery rate, especially for weak signals.

IP14 — Recent Developments in Statistical Machine Learning and Causal Inference

June 14 (Sunday) 13:00–14:40 **Room:** 209A Program

Quefeng Li, The University of North Carolina at Chapel Hill, United States

Inference on the Significance of Modalities in Multimodal Generalized Linear Models

Multimodal statistical models have gained attention in recent years, yet there lacks rigorous statistical inference tools for inferring the significance of a single modality within a multimodal model. This inference problem is particularly challenging in high-dimensional multimodal models. In the context of high-dimensional multimodal generalized linear models, we propose a novel entropy-based metric, called the expected relative entropy, to quantify the information gain of one modality in addition to all other modalities in the model. We develop a deviance-based

statistic to estimate the expected relative entropy and prove that this statistic is consistent and show that its asymptotic distribution can be approximated by a non-central chi-squared distribution. That enables the calculation of confidence intervals and p-values to assess the significance of the expected relative entropy for a given modality. Such an inference tool is useful for ranking the importance of data modalities and making personalized treatment recommendations based on individual patient profiles.

Zhengling Qi, The George Washington University, United States

InSPO: Unlocking Implicit Self-Reflection for LLM Preference Optimization

Direct Preference Optimization (DPO) and its variants have become the standard for aligning Large Language Models (LLMs) due to their simplicity and offline stability. However, we identify two fundamental limitations that undermine the reliability and optimality of DPO and its equivalence (i.e., reinforcement learning with human feedback (RLHF)). First, they lack invariance: the optimal policy derived by current methods changes based on arbitrary modeling choices such as the scalarization function (e.g, logistic function in Bradley-Terry model) or reference policy, which may yield behavior that is an artifact of parameterization rather than a reflection of true human preference. Second, they are theoretically sub-optimal because they treat response generation in isolation, failing to leverage the comparative information embedded in the pairwise preference data. This restricts the model's inherent ability to "compare and contrast" responses, leaving its capacity for implicit self-reflection untapped. In this work, we propose a novel family of Implicit Self-reflective Preference Optimization (InSPO) methods, which address these issues. We first derive a globally optimal policy that should condition on both the context and the alternative response under the pairwise preference data setting, which unleashes the self-reflection property explicitly. Then we theoretically demonstrate that this formulation is superior to standard DPO and RLHF targets and guarantees invariance to the choice of scalarization and reference policy. Practically, InSPO operationalizes this target as a plug-and-play enhancement for DPO-family algorithms, decoupling the alignment goal from modeling constraints without requiring complex architectural changes. Crucially, our method incurs zero inference overhead, as the self-reflective mechanism is distilled into the policy during training and does not require generating alternative responses during deployment. Comprehensive experiments demonstrate that InSPO achieves consistent improvements in win rates and length-controlled metrics, validating that unlocking self-reflection leads to more robust and human-aligned LLMs.

Will Wei Sun, Purdue University, United States

When Is an LLM Really Better? Uncertainty-Aware Evaluation from Sparse Human Preferences

Large language model evaluation increasingly relies on pairwise human preferences, as in Arena-style platforms, but such data are often sparse, imbalanced, and heterogeneous across tasks. As a result, point-valued leaderboards can be unstable and may overstate small performance differences. In this talk, I will present statistical frameworks for uncertainty-aware LLM evaluation from sparse pairwise comparisons. We formulate LLM evaluation as a low-rank completion problem, using a low-rank task-by-model ability structure to share information across related tasks while preserving task-specific heterogeneity. This framework enables semiparametrically efficient inference for model ability gaps and win probabilities, and further provides entrywise estimation guarantees, top-K recovery theory, debiased score-gap inference, and simultaneous confidence sets for ranks and top-K membership.

Gongjun Xu, University of Michigan, United States

Identifiability and Inference for Generalized Latent Factor Models

Generalized latent factor analysis not only provides a useful latent embedding approach in statistics and machine learning, but also serves as a widely used tool across various scientific fields, such as psychometrics, econometrics, and social sciences. Ensuring the identifiability of latent factors and the loading matrix is essential for the model's estimability and interpretability, and various identifiability conditions have been employed by practitioners. However, fundamental statistical inference issues for latent factors and factor loadings under commonly used identifiability conditions remain largely unaddressed, especially for correlated factors and/or non-orthogonal loading matrix. In this work, we focus on the maximum likelihood estimation for generalized factor models and establish statistical inference properties under popularly used identifiability

conditions. The developed theory is further illustrated through numerical simulations and an application to a personality assessment dataset.

IP18 — Inference under Heterogeneity and Side Information

June 14 (Sunday) 13:00–14:40 **Room:** 209B Program

Kabir Verchand, University of Southern California, United States

Statistical-computational gaps in estimation with missing not at random data

Estimators designed to handle missingness often rely on strong assumptions about the mechanism by which data is missing, such as that the data is missing completely at random (MCAR). By contrast, real data is rarely MCAR. In the absence of these strong assumptions, can we still trust these estimators? Towards understanding this question, I will introduce a robust framework which bridges the gap between the MCAR and assumption-free settings. Focusing on the fundamental task of mean estimation, I will demonstrate that even in the univariate setting, there is an intricate tradeoff between estimation accuracy and robustness to modeling assumption. Finally, I will turn to the higher-dimensional setting and provide evidence—through a statistical query lower bound—towards a statistical–computational gap in this robust setting as well as an efficient sum-of-squares based algorithm which saturates this statistical query lower bound.

Oh-Ran Kwon, The Ohio State University, United States

Black-Box Knowledge Transfer for Distinct Feature Sets

Pre-trained black-box predictive functions encode valuable knowledge, distilled from massive datasets and extensive computation. However, when the available input features come from an input space that differs from that of the black box, direct use is infeasible. A natural approach is to apply the black box through a mapping between the two input spaces, but this breaks down when the black box is highly nonlinear. Instead, in this talk, we introduce a method for transferring predictive knowledge from the black box to a different input space. Our approach decomposes the target prediction function into two components: a transferable component, which can be informed by the black box, and a non-transferable component, which captures information unique to the new space. We introduce a two-step estimation procedure aligned with this decomposition. We derive non-asymptotic prediction error bounds and show that transfer learning is advantageous over a non-transfer alternative, particularly when the non-transferable component is small or smooth. We further extend our approach to the case where multiple black-box functions are available and show that aggregating them provably improves predictive performance. Simulated and real data examples demonstrate the practical value of the proposed approach.

Fei Xue, Purdue University, United States

An Empirical Bayes Regression for Multi-tissue Gene Expression Prediction

The Genotype-Tissue Expression (GTEx) project collects samples from multiple human tissues to study the relationship between genetic variation or single nucleotide polymorphisms (SNPs) and gene expression in each tissue. However, most existing eQTL analyses only focus on single tissue information. In this paper, we develop a multi-tissue method that improves prediction of gene expression based on cis-SNPs by borrowing information across tissues. Specifically, we propose an empirical Bayes regression model for SNP-expression association using data from multiple tissues. To allow the effects of SNPs to vary greatly among tissues, we use a mixture distribution as the prior, which is a mixture of a multivariate Gaussian distribution and a Dirac mass at zero. We show that the proposed estimator of the cis-SNP effects on gene expression asymptotically achieves the minimum Bayes risk among all estimators. Analyses of the GTEx data show that our proposed method is superior to existing methods in terms of prediction accuracy for gene expression using cis-SNPs in testing sets.

Keisuke Yano, The Institute of Statistical Mathematics, Japan

Inference under Frequency-Domain Heterogeneity via Spectral Rényi Divergences

In this talk, we discuss inference under frequency-domain heterogeneity, by which we mean heterogeneity in the reliability and informativeness of different frequency components in time series data. We study a specific class of statistical divergences for spectral densities of time series: spectral alpha-Rényi divergences, which include the Itakura–Saito divergence as a limiting case. The goal of this talk is to highlight both information-theoretic and statistical properties of spectral alpha-Rényi divergences. We reveal a connection between spectral alpha-Rényi divergence and gamma-divergence in robust statistics, and establish a variational representation of spectral alpha-Rényi divergence. Motivated by these results, which suggest a form of robustness, we show that the minimum spectral Rényi divergence estimator has a stable optimization path in the presence of outliers in the frequency domain. This is in contrast to the minimum Itakura–Saito divergence estimator. Consequently, the proposed approach can deliver more stable estimates and reduce the need for delicate preprocessing. We also illustrate the proposed approach through an application to geodetic data.

IP28 — Memory and Learning in Probability

June 14 (Sunday) 13:00–14:40 **Room:** 203 Program

Shuhei Mano, The Institute of Statistical Mathematics, Japan

Symmetric quantum walks on Hamming graphs and their limit distributions

We study a class of symmetric coined quantum walks on Hamming graphs, where the distance between vertices specifies the transition probability. A special model is the simple quantum walk on the hypercube, which has been discussed in the literature. Eigenvalues of the unitary operator of the quantum walks are zeros of certain self-reciprocal polynomials. We obtain a spectral representation of the wave vector, where our systematic treatment relies on the coin space isomorphic to the state space and the commutative association scheme. The Grover coin is extended to the reflection about the invariant subspace of the Terwilliger algebra. The limit distributions of several quantum walks are obtained. This is a joint work with Robert Griffiths.

Anish Sarkar, Indian Statistical Institute, New Delhi, India

Scaling limit of a drainage network model on perturbed lattice

Study of random networks originating from nodes which are either fixed or distributed independently and uniformly over space have been studied in the drainage network models and their related versions. These includes points retained with a fixed probability on the standard $\mathbb{Z}^2$ lattice or the Poisson point process in the continuum set up. In this work, we venture beyond this standard paradigm and investigate a stochastic forest obtained from a drainage network model constructed on a randomly perturbed subset of $\mathbb{Z}^2$, where both horizontal and vertical perturbations are given by exponentially decaying unbounded discrete random variables and vertical perturbations are allowed in the upward direction only. We show that the resultant stochastic network is a single tree a.s. We further establish that as a collection of paths, under diffusive scaling the resultant network converges to the Brownian web

Neeraja Sahasrabudhe, Indian Institute of Science Education and Research, Mohali, India

Elephant Random Walk with Tampered Memory

TBD

Moumanti Podder, Indian Institute of Science Education and Research, Pune, India

A model of market economics inspired by random walks with long memory

Imagine maintaining a chart for tracking the relative performance of one product that competes against another in an oligopolistic market. Every time an item of the first product is purchased, we add a +1 to our chart, and every time an item of the second product is purchased, we add a -1 to our chart, giving rise to a random-walk-like model where the relative performance of the first product with respect to the second, after a total of n items of these two products have been

bought, is the same as the position of the walker that begins walking from the origin. In scenarios where customers may not have complete information regarding the underlying qualities of these two products, a customer may draw a sample of k customers from the past, and note the choices that were made by these sampled customers. Based on these sampled observations, the customer decides, according to a stochastic rule, which product to choose. This is reminiscent of a random walk where the walker decides every step based on their memory of the entire past, and can be thought of as a generalized version of the relatively new and extremely popular random walk model known as the 'elephant random walk'. In this talk, I discuss results pertaining to strong and weak convergence of this generalized elephant random walk.

IP38 — Mixing Behavior of Markovian Dynamics

June 14 (Sunday) 13:00–14:40 **Room:** 201 Program

Seonwoo Kim, Yonsei University, South Korea

Transience time of the subcritical facilitated exclusion process

In this talk, we consider the facilitated exclusion process on the one-dimensional discrete N-torus. Because of the facilitating mechanism, the process freezes in finite time if the particle density of the initial configuration is subcritical, i.e., if it is smaller than (or equal to) $1/2$. We prove that, starting from any subcritical Bernoulli product measure, the correct scale of the transience/freezing time is of order $\log^3 N$. In addition, if the density is near criticality, we show that the transience time undergoes another phase transition at $\alpha = 1/2$ where the density is $1/2 - N^{-\alpha}$. Based on a joint work with Oriane Blondel, Clément Erignoux and Sanha Lee.

Eric O. Endo, New York University Shanghai, China

Phase Transition of the Long-Range Ising Models with Cell-Board External Fields

On the two-dimensional lattice $\mathbb{Z}^2$, the phase transition of the nearest-neighbor ferromagnetic Ising model with periodical external fields, which takes two values h and $-h$, where $h > 0$, has been extensively studied in the literature. González-Navarrete, Pechersky, and Yambartsev (J. Stat. Phys.-2016) considered a more general setting by associating to each site a positive and negative values of external field arranged in a cell-board configuration with rectangular cells of size $L_1 \times L_2$ sites, such that the total value of the external field vanishes. They proved that, for the nearest-neighbor model, a phase transition occurs at low temperatures for all $h < 2J/L_1 + 2J/L_2$, where $J > 0$ is the coupling constant.

In this talk, we extend this setting to long-range ferromagnetic interactions of the form $J_{xy} = \|x - y\|^{-\alpha}$, with $\alpha > 2$. We derive conditions on h for the existence of a phase transition at low temperatures in the presence of cell-board external fields.

Joint work with Yuliang Shi (UBC-Canada).

Chiara Franceschini, Università di Modena e Reggio Emilia, Italy

Degree-preserving conservative processes and a unified approach for their hydrodynamics

In this talk we will consider a broad class of interacting systems characterized by a single conservation law and satisfying the "degree-preserving property":

namely, the Markov generator of the processes preserves the degree of polynomials of the state variables up to two

This, together with other mild assumption, allows us to propose a new universal framework to rigorously derive the hydrodynamic limit of the class of processes without considering the peculiarity of each models. These models admit stationary measures of product form whose marginals are given by only six different densities, yielding a complete classification consistent with the natural exponential family setting of Morris [1]. As a consequence, we construct a new model whose invariant measure is given by the generalized hyperbolic secant distribution. This work [2] is in collaboration with Patrìcia Gonçalves, Kohei Hayashi, Makiko Sasada and generalizes our previous result [3].

[1] C. N. Morris. Natural exponential families with quadratic variance functions. The Annals of Statistics, pages 65–80, 1982.

[2] C. Franceschini, P. Gonçalves, K. Hayashi, and M. Sasada. Classification of product invariant measures for degree-preserving conservative processes and their hydrodynamics. arXiv preprint arXiv:2604.03548.

[3] C. Franceschini, P. Gonçalves, K. Hayashi, and M. Sasada. Hydrodynamic limit for some gradient and attractive spin models. Journal of Statistical Physics, 193(1):3, 2026.

Jungkyoung Lee, Inha University, South Korea

Mixing of the Curie–Weiss–Potts model

Mixing refers to the convergence of the distribution of a stochastic dynamical system to its stationary distribution. In the high-temperature regime, the Glauber dynamics of the Curie–Weiss–Potts model exhibits fast mixing, which arises from the concentration of the stationary distribution around a unique equilibrium. In contrast, in the low-temperature regime, the presence of multiple equilibria leads to slow mixing, since transitions between distinct equilibria are required for the system to reach stationarity. In this talk, we discuss the slow mixing of the Glauber dynamics for the Curie–Weiss–Potts model in the low-temperature regime.

This talk is based on joint with Seonwoo Kim (Yonsei University).

IP36 — Recent Advances in Nonparametric Bayesian Learning

June 14 (Sunday) 13:00–14:40 **Room:** 202 Program

Minwoo Chae, Pohang University of Science and Technology, South Korea

Nonparametric Estimation of Undirected Graph Structures Using Diffusion Models

An undirected graph represents the unconditional independence structure among random variables. Although estimating undirected graphs is important in many applications, most existing methods are restricted to parametric settings such as Gaussian graphical models and Ising models. In this talk, we consider a nonparametric approach to undirected graph estimation based on diffusion models. Recent work has shown that density estimators based on diffusion models can adapt to the unknown undirected graph structure of the data; however, diffusion models themselves do not provide an explicit estimator of the graph. To address this limitation, we propose a novel method for undirected graph estimation that does not rely on parametric assumptions.

Weining Shen, University of California, Irvine, United States

Deep kernel learning based Gaussian processes for Bayesian image regression

Regression models are widely applied in neuroimaging studies to learn associations between clinical and image variables. Gaussian process (GP) priors are common Bayesian nonparametric approaches to model unknown regression functions in high-dimensional data. However, existing GP methods depend on pre-specified parametric covariance kernels, which often lack the flexibility to capture data complexity, have limited generalizability across study populations, and face computational challenges in large-scale datasets. We propose a scalable fully Bayesian deep kernel learning framework for GP priors with applications in various image regression models. Our method leverages the estimation power of deep neural networks (DNNs) to adaptively learn the kernel basis functions from the data to capture complex spatial correlations. We establish theoretical properties by deriving posterior concentration rates of regression and kernel function estimation. Simulation studies demonstrate improved accuracy in estimation and signal detection across different scenarios. We further validate the proposed method through two neuroimaging applications.

Seonghyun Jeong, Yonsei University, South Korea

Bayesian Spatially Adaptive Triangulation

Conventional nonparametric regression on two-dimensional non-rectangular domains often overlooks domain geometry, allowing smoothing across boundaries. In spatial and geostatistical applications, this assumption is frequently invalid, as domain boundaries typically constrain interactions among observations. Addressing spatially varying smoothness is also substantially more difficult than in the univariate setting, and most existing methods fail to capture this local property of the target function. To address these challenges, we propose Bayesian spatially adaptive triangulation, which constructs

locally adaptive splines over a polygonal domain. The method employs constrained Delaunay triangulations optimized to represent boundary geometry and facilitate adaptation to varying smoothness levels. A carefully designed prior enhances empirical performance. Under a global smoothness assumption, we show that the proposed method achieves the optimal posterior contraction rate and adapts to unknown smoothness. We also show that the method exhibits ideal spatial adaptation in the sense that it achieves the oracle rate for inhomogeneous or locally varying structural features. Simulation studies confirm that it outperforms existing approaches, achieving higher estimation accuracy while maintaining low model complexity.

Gyuhyeong Goh, Kyungpook National U, South Korea

Bayesian infinite interactive fixed effects modeling for causal inference

Causal inference for single treatment effect estimation is challenging due to the absence of valid control units. The synthetic control method (SCM) offers an innovative way of constructing the so-called data-driven control unit. The generalized synthetic control (GSC) method is proposed as a factor model-based extension of SCM. While GSC improves upon SCM, the performance of GSC heavily depends on the choice of the number of latent factors. To account for the uncertainty associated with the number of factors, we propose to employ a Bayesian infinity factor modeling approach. The key idea of our Bayesian infinity factor modeling is to assign a cumulative shrinkage process prior on the factor loadings. In addition, we apply a Gaussian process approach to infer the non-linear treatment effect. The proposed Bayesian framework enables us to make full Bayesian inference about the time-varying treatment effect. The merits of the proposed Bayesian method are demonstrated through simulation studies and real data analysis.

IP07 — Advances in Change Point Detection and Dependence Structures in High Dimensions

June 14 (Sunday) 13:00–14:40 **Room:** 214 Program

Baojun Dou, City University of Hong Kong, Hong Kong

Pairs Trading: Cointegration or Lack of It

This paper develops a novel framework for online detection of regime changes in pairs trading relationships. We introduce a two-stage methodology where the first stage estimates the cointegration relationship to construct the trading spread, and the second stage monitors this spread for changes in mean-reversion properties. Our approach addresses the practical challenge that cointegration relationships are not permanent but evolve over time due to market regime changes. We develop sequential monitoring procedures that provide actionable signals for dynamic position management, allowing traders to reduce exposure during spread deterioration and build positions during spread emergence. Through theoretical analysis, simulations, and empirical backtesting on the EWA-EWC ETF pair, we demonstrate that our approach leads to improved risk management and better trading performance compared to strategies without regime change detection.

Zetai Cen, University of Bristol, United Kingdom

Change Point Detection and Identification in Tensor Factor Models

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.

Yue Du, Southwestern University of Finance and Economics, China

IDENTIFICATION AND ESTIMATION FOR MATRIX TIME SERIES CP-FACTOR MODELS

We propose a new method for identifying and estimating the CP-factor models for matrix time series. Unlike the generalized eigenanalysis-based method of Chang et al. (2023) for which the convergence rates of the associated estimators may suffer from small eigengaps as the asymptotic theory is based on some matrix perturbation analysis, the proposed new method enjoys faster convergence rates which are free from any eigengaps. It achieves this by turning the problem into a joint diagonalization of several matrices whose elements are determined by a basis of a linear system, and by choosing the basis carefully to avoid near co-linearity .

Furthermore, unlike Chang et al. (2023) which requires the two factor loading matrices to be full-ranked, the proposed new method can handle rank-deficient factor loading matrices. Illustration with both simulated and real matrix time series data shows the advantages of the proposed new method.

Shakeel Gavioli-Akilagun, City University of Hong Kong, Hong Kong

Optimal Online Change Detection via Random Fourier Features

We study the problem of online non-parametric change point detection in multivariate data streams. We approach the problem through the lens of kernel-based two-sample testing and introduce a sequential testing procedure based on random Fourier features, running with logarithmic time complexity per observation and with overall logarithmic space complexity. The algorithm has two advantages compared to the state of the art. First, our approach is genuinely online, and no access to training data known to be from the pre-change distribution is necessary. Second, the algorithm does not require the user to specify a window parameter over which local tests are to be calculated. We prove strong theoretical guarantees on the algorithm's performance, including information-theoretic bounds demonstrating that the detection delay is essentially optimal in the minimax sense. Numerical studies on real and synthetic data show that our algorithm is competitive with respect to the state of the art.

This is joint work with Florian Kalinke (KIT) and the full paper can be found here: https://neurips.cc/virtual/2025/loc/san-diego/poster/118949.

CS07 — Nonparametric and Semiparametric Statistical Methods

June 14 (Sunday) 13:00–14:40 **Room:** 215 Program

Dennis Leung, University of Melbourne

Berry-Esseen Theorems for the Asymptotic Normality of Incomplete U-Statistics with Bernoulli Sampling

There has been a resurgence of interest in incomplete U-statistics that only sum over a subset of kernel evaluations, due to their computational efficiency and asymptotic normality which can be leveraged to quantify the uncertainty of ensemble predictions in machine learning. I will revisit the weak convergences to normality of one such construction, the incomplete U-statistic with Bernoulli sampling, under three different regimes on the relative sizes of the raw sample and the computational budget that trace back to the work of Janson (1984). Specifically, I will talk about my recently established Berry-Esseen bounds with the natural rates that characterize the accuracy of these normal approximations, which were published in Leung (2026, Electronic Journal of Probability, 31, 1-52). I will also discuss some of my follow-up work in progress.

Taishi Kuzumoto, University of Tokyo

Inference on Locally Adaptive Nonparametric Regression

In nonparametric regression, classical linear smoothers such as kernels and splines have enabled the construction of statistical inference methods, including confidence bands, by exploiting their linearity when the underlying function class has globally uniform smoothness. However, such linear smoothers are known to lack local adaptivity; that is, they cannot accommodate varying degrees of smoothness of the function across the domain.

On the other hand, as an example of a locally adaptive nonparametric regression method, trend filtering has been proposed. Due to its nonlinearity, however, it has been difficult to develop statistical inference procedures using conventional probabilistic tools. In this study, we derive an identification representation that captures the local behavior of the trend filtering estimator, and furthermore, based on the idea of resampling methods, we approximate the distribution of unknown and unobservable statistics to propose a method for constructing confidence bands.

Deborshi Das, Indian Statistical Institute, Delhi

Elephant Random Walks with Graph Based Shared Memory

We introduce a generalized model of the elephant random walk, featuring multiple elephants moving along the integer line, Z, and interacting through a shared memory structure governed by a directed graph. In this framework, each elephant' s next step depends not only on its own past trajectory but also on the past steps of other elephants, based on the graph structure. Each vertex in the graph represents an elephant, and directed edges indicate that an elephant considers the previous steps of its in-neighbors when determining its next move. This model thus represents a system of reinforced random walks, evolving under graph-based interdependencies. The first- and second-order asymptotic behavior of the joint walks will be briefly covered, further an outline of the proof techniques and connection to other network-based reinforced processes will also be covered.

Daisuke Matsuno, Tohoku University Graduate School of Information Sciences

Semiparametric Regression with Stagewise Minimization of the Empirical Risk

This study addresses multiple regression problems via a semiparametric approach combining interpretable parametric estimation and its flexible nonparametric adjustment.

In the proposed estimator, initial estimation is made by a parametric approach such as generalized linear regression. This approach models the target regression function as functions determined by a finite-dimensional parameter, offering interpretations for the target one. In generalized linear models, for example, the parameters correspond to coefficients of a hyperplane, and allow us to get information: whether the regression function increases, decreases, or hardly changes globally with some explanatory variables.

On the other hand, such parametric models are sometimes so restrictive that estimators obtained by these models can not be consistent in general. Then without losing the interpretability via the parametric approach, the flexible nonparametric one is also utilized to adjust the parametric crude guess and make the estimator consistent.

The nonparametric adjustment in the proposed estimator is realized as a convex combination of several words (functions) selected from a prepared dictionary (set of functions). This convex combination is constructed through stagewise minimization of the empirical risk.

A non-asymptotic error bound for this proposed algorithmic estimator is established. We also propose an effective dictionary associated with a variant of Nadaraya-Watson estimator, and report results of Monte Carlo simulation for the concrete proposed estimator.

Eun-Ji Lee, Chungbuk National University

Simultaneous Estimation of Nonparametric Quantile Regression with B-Splines and Norm-Based Penalties

In this study, we propose a B-spline quantile regression estimator that incorporates group-norm and nuclear-norm penalties. The group-norm penalty promotes shared selection of spline knots across quantile functions, encouraging a common functional structure and improving interpretability. The nuclear-norm penalty enables simultaneous estimation of multiple quantiles, ensuring structural coherence and enforcing non-crossing of the estimated curves. The estimator is implemented via the alternating direction method of multipliers, with an adaptive penalty parameter update scheme to improve convergence stability and estimation accuracy across varying sample sizes and signal-to-noise ratios. Together, these complementary penalties facilitate smooth and well-structured estimation while preserving meaningful differences across the response distribution. Its performance is evaluated through simulations considering constant, linear, and cubic functional forms under both homoscedastic and heteroscedastic settings. Application to empirical data demonstrates the practical utility of the approach and the benefits of the adaptive penalty adjustment.

DL06 — To be confirmed

June 14 (Sunday) 14:50–16:30 **Room:** LT1A Program

Jun Liu, Harvard University, United States

Conditional Generation via Diffusion, Flow, and Schrodinger Bridges

Generative modeling has become a popular topic and powerful tool in AI. Conditional generation (or sampling) refers to tasks of generating targets that met certain prespecified conditions. In the training-free setting, a pretrained unconditional model is reused as a prior and combined with a task-specific likelihood to sample from the corresponding posterior, without retraining the generative model. Recently, sequential Monte Carlo (SMC, aka particle filtering) techniques have been employed for such tasks via both diffusion and flow-based generative models. We will describe some of these advances. Furthermore, training-free conditional sampling for Schrödinger bridge (SB) generative models remains unexplored, despite the growing appeal of SBs as flexible stochastic transport models between general source and target distributions. We will describe BridgeTwist, a training-free conditional sampler for pretrained SB generative models, formulated as a tempered twisted SMC scheme. Two key steps in SMC are (a) a good sampling distribution (or a good way to "twist" the particle generation); and (b) an effective resampling scheme, whose role is to guide resources towards promising particles. We will review some approaches for conducting these two main steps in both diffusion, flow, and SB models. We also propose Wasserstein-Dirichlet resampling (WDR). WDR first constructs an empirical measure that optimally approximates the weighted particle distribution in the Wasserstein sense by solving a free-support Wasserstein barycenter problem. To balance geometric fidelity with Monte Carlo variability, WDR further employs a Dirichlet mixing mechanism that randomizes the optimal coupling. This talk is based on the joint work with Qianqian Qu, Mengyu Li and Cheng Meng.

Ke Deng, Tsinghua University, China

Adaptive Kernel Density Estimation with Pre-training

Density estimation in high-dimension is an important statistical problem with great challenges. Traditional methods based on kernel smoothing are inefficient in high-dimension due to the difficulties in specifying appropriate location-adaptive kernels. In this talk, we introduce pre-training, a key idea behind many cutting-edge AI technologies, to the context of nonparametric density estimation. By establishing a pre-trained neural network that can recommend an appropriate location-adaptive kernel for each sample point, efficient density estimation with adaptive kernels is achieved in high-dimension. A wide range of numerical experiments show that this strategy is highly effective to improve the accuracy of density estimation, when the target distribution belongs to the distribution family for pre-training. When the target distribution is far way from the distribution family for pre-training, the benefit from the proposed pre-training strategy may be diluted, but can be reactivated by an extra fine-tuning procedure.

Minsuk Shin, Yonsei University, South Korea

Amortizing Bootstrapped Nonparametric Maximum Likelihood Estimator

Nonparametric maximum likelihood estimation (NPMLE) is a central tool for mixture models, but bootstrap-based inference is computationally demanding because each bootstrap replicate requires solving an infinite-dimensional optimization problem from scratch. We propose GB-NPMLE, a generative bootstrap framework that recasts bootstrap inference as amortized posterior-mass inference. Rather than re-solving the NPMLE for every bootstrap draw, we learn a single conditional map that takes a noise scale and a multinomial bootstrap weight vector as input and returns a discrete probability law over a collection of learnable support points, reducing each bootstrap draw to a single forward pass. The map is parameterized by a weight-modulated MLP that fuses scale and bootstrap-weight information through additive and FiLM-style conditioning, refined by residual blocks, and trained to maximize a numerically stabilized bootstrap-weighted marginal log-likelihood. On the theoretical side, we establish asymptotic normality of the NPMLE plug-in and consistency of the multinomial bootstrap for smooth linear functionals of the mixing distribution, using a bracketing-entropy Donsker property of the mixture class and its score space. We also clarify non-regular regimes in which the standard bootstrap fails and amortization should not be expected to recover the correct limit without modification.

DL04 — Minimax Optimality in Online Learning

June 14 (Sunday) 14:50–16:30 **Room:** LT1B Program

Alexandre B. Tsybakov, ENSAE, France

Continuum bandit, gradient-free stochastic optimization and nonparametric regression

This talk will consider continuum (contextual) bandit and gradient-free stochastic optimization problems through the lens of nonparametric regression. We will describe the similarities and differences, and further focus on deriving information theoretical limits and algorithms that are optimal in a minimax sense for these sequential decision making settings.

Arya Akhavan, University of Oxford, United Kingdom

Non-stationary bandit convex optimization: A comprehensive study

Bandit Convex Optimization is a fundamental class of sequential decision-making problems, where the learner selects actions from a continuous domain and observes a loss (but not its gradient) at only one point per round. We study this problem in non-stationary environments, and aim to minimize the regret under three standard measures of non-stationarity: the number of switches in the comparator sequence, the total variation of the loss functions, and the path-length of the comparator sequence. We propose a polynomial-time algorithm, Tilted Exponentially Weighted Average with Sleeping Experts (TEWA-SE), which adapts the sleeping experts framework from online convex optimization to the bandit setting. For strongly convex losses, we prove that TEWA-SE is minimax-optimal with respect to known and by establishing matching upper and lower bounds. By equipping TEWA-SE with the Bandit-over-Bandit framework, we extend our analysis to environments with unknown non-stationarity measures. For general convex losses, we introduce a second algorithm, clipped Exploration by Optimization (cExO), based on exponential weights over a discretized action space. While not polynomial-time computable, this method achieves minimax-optimal regret with respect to known and , and improves on the best existing bounds with respect to .

Alexandra Carpentier, University of Potsdam, Germany

A simple and improved algorithm for noisy, convex, zeroth-order optimisation

We consider the problem of zeroth-order noisy convex optimisation. In this setting, we can observe sequentially and adaptively noisy evaluations of a convex function f defined on R^d and taking value in R. Namely, we can choose the sampling points depending on past observations. The aim is to find a near-minimum of this function over a given compact, convex set, given a budget of n evaluations. While this problem is simple and important, it is quite challenging and many important questions remain open. In this talk, I will discuss a simple method proposed in [1] for solving it, based on an adaptation of the barycenter method. An important note is that some more powerful alternative methods exist, see [2]. However, none of them is proven to be optimal in the worst case.

[1] Carpentier, A. (2025). A simple and improved algorithm for noisy, convex, zeroth-order optimisation. Mathematical Statistics and Learning, 8(3), 165-192. [2] Fokkema, H., van der Hoeven, D., Lattimore, T., & Mayo, J. J. (2024). Online Newton method for bandit convex optimisation. arXiv preprint arXiv:2406.06506.

IP01 — Recent Development on High-Dimensional Data Modeling

June 14 (Sunday) 14:50–16:30 **Room:** 209A Program

Yingying Li, The Hong Kong University of Science and Technology, Hong Kong

Site Percolation Network Models for Event-Driven Systems

We propose a class of network models for event-driven systems based on site percolation, where edges arise from the simultaneous activation of nodes. Unlike standard edge-based models, our framework captures dependence induced by

common shocks and naturally represents transitivity in co-activation networks. We develop spectral estimators that consistently recover community structure under both pure and mixed membership specifications. We further show that these estimators admit a GMM interpretation based on second-moment conditions. We demonstrate the model's practical relevance by recovering latent structures in macroeconomic shocks, firm characteristic co-movements, and co-citation networks. Based on joint work with Xinghua Zheng.

Jingyuan Liu, Xiamen University, China

LLM-Powered Deep Panel Modeling with Application to Regional CPI Prediction

Understanding regional Consumer Price Index (CPI) dynamics is essential for timely and effective economic policymaking. However, traditional modeling procedures typically rely only on low-frequency, high-cost surveys and macroeconomic indices, which often fail to capture rapid market fluctuations and lead to inaccurate predictions. To address such challenges, we propose a new framework that integrates large language model (LLM) analyses and social media narratives with dynamic panel modeling. We construct a narrative corpus from a newly-collected Sina Weibo dataset, and introduce a prompt-based GPT model and a series of fine-tuned BERT models to generate high-frequency LLM-induced surrogates for regional CPI. A novel joint modeling strategy is advocated to transfer the information from these surrogates to the targeted official CPI data and hence empower CPI prediction. To solve the joint objectives, we further introduce a new deep panel learning procedure with region-wise homogeneity pursuit, which has its own significance in panel data analysis literature. In addition, conformal-based panel prediction intervals are provided to quantify the uncertainty of the LLM-powered prediction. Empirical and theoretical analyses demonstrate that our approach significantly reduces short-term forecasting errors and more effectively captures abrupt inflationary shifts compared to traditional econometric models. While demonstrated for regional CPI prediction, the proposed framework is broadly applicable for incorporating insights from LLMs to enhance traditional statistical modeling.

Zhanrui Cai, The University of Hong Kong, Hong Kong

A Statistical Framework for Alignment with Biased AI Feedback

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.

Yei Eun Shin, Seoul National University, South Korea

Dynamic Network Modeling for the Spatiotemporal Progression of High-Dimensional Data

The modeling of high-dimensional spatiotemporal data is essential for understanding complex dynamic systems, such as tracking the large-scale spreading patterns of amyotrophic lateral sclerosis (ALS) across a human body map. This presentation reviews various methodological developments designed to model the progression of these network dependencies over time without relying on prespecified spatial proximity. We first highlight an autologistic network model for binary data with absorbing states. This approach employs sparse regularization to effectively manage a large number of pairwise spatial associations. We then introduce a spatial hidden Markov model, which accounts for diagnostic misclassifications through high-dimensional inference of latent disease states. Expanding the discussion to broader modeling strategies, we also introduce our ongoing research that handles longitudinal observations at their original continuous level. In this framework, complex high-dimensional network dependencies are flexibly captured using a linear mixed-effects model by specifying the covariance structure. Specifically, this is achieved based on inter-node distances measured by the resistance metric, a

type of non-Euclidean metric. Alongside these diverse approaches, additional ongoing extensions and related follow-up studies will be discussed to explore the broader scalability and potential of high-dimensional network modeling.

IP19 — Recent Advances in High-dimensional Statistics

June 14 (Sunday) 14:50–16:30 **Room:** 209B Program

Dong Xia, The Hong Kong University of Science and Technology, Hong Kong

TBD

TBD

Peter Radchenko, The University of Sydney, Australia

Extracting Interpretable Models from Tree Ensembles

Tree ensembles are non-parametric methods widely recognized for their accuracy and ability to capture complex interactions. While these models excel at prediction, they are difficult to interpret and may fail to uncover useful relationships in the data. We propose an estimator to extract compact sets of decision rules from tree ensembles. The extracted models are accurate and can be manually examined to reveal relationships between the predictors and the response. A key novelty of our estimator is the flexibility to jointly control the number of rules extracted and the interaction depth of each rule, which improves accuracy. We develop an exact algorithm to efficiently solve optimization problems underlying our estimator and an approximate algorithm for computing regularization paths —sequences of solutions that correspond to varying model sizes. We also establish novel non-asymptotic prediction error bounds for our proposed approach, comparing it to an oracle that chooses the best data-dependent linear combination of the rules in the ensemble subject to the same complexity constraint as our estimator. Through experiments, we demonstrate that our estimator outperforms existing algorithms for rule extraction.

Anderson Ye Zhang, University of Pennsylvania, United States

Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery

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

Guillaume Braun, RIKEN, Japan

Learning Dynamics of Phase Retrieval under Power-Law Data

Scaling laws describe how learning performance improves with data, compute, or training time, and have become a central theme in modern deep learning. We study this phenomenon in a canonical nonlinear model: phase retrieval with anisotropic Gaussian inputs whose covariance spectrum follows a power law. Unlike the isotropic case, where dynamics collapse to a two-dimensional system, anisotropy yields a qualitatively new regime in which an infinite hierarchy of

coupled equations governs the evolution of the summary statistics. We develop a tractable reduction that reveals a three-phase trajectory: (i) fast escape from low alignment, (ii) slow convergence of the summary statistics, and (iii) spectral-tail learning in low-variance directions. From this decomposition, we derive explicit scaling laws for the mean-squared error, showing how spectral decay dictates convergence times and error curves. Experiments confirm the predicted phases and exponents. These results provide the first rigorous characterization of scaling laws in nonlinear regression with anisotropic data, highlighting how anisotropy reshapes learning dynamics.

IP22 — Statistical Modeling with Deep Learning and Biomedical Applications

June 14 (Sunday) 14:50–16:30 **Room:** 203 Program

Lican Kang, Wuhan University, China

TBD

TBD

Ziyuan Chen, Peking University, China

Deep Semiparametric Partial Differential Equation Models

In many scientific fields, the generation and evolution of data are governed by partial differential equations (PDEs) which are typically informed by established physical laws at the macroscopic level to describe general and predictable dynamics. However, some complex influences may not be fully captured by these laws at the microscopic level due to limited scientific understanding. This work proposes a unified framework to model, estimate, and infer the mechanisms underlying data dynamics. We introduce a general semiparametric PDE (SemiPDE) model that combines interpretable mechanisms based on physical laws with flexible data-driven components to account for unknown effects. The physical mechanisms enhance the SemiPDE model's stability and interpretability, while the data-driven components improve adaptivity to complex real-world scenarios. A deep profiling M-estimation approach is proposed to decouple the solutions of PDEs in the estimation procedure, leveraging both the accuracy of numerical methods for solving PDEs and the expressive power of neural networks. For the first time, we establish a semiparametric inference method and theory for deep M-estimation, considering both training dynamics and complex PDE models. We analyze how the PDE structure affects the convergence rate of the nonparametric estimator, and consequently, the parametric efficiency and inference procedure enable the identification of interpretable mechanisms governing data dynamics. Simulated and real-world examples demonstrate the effectiveness of the proposed methodology and support the theoretical findings.

Daewoo Pak, Yonsei University, South Korea

Genome-Wide Identification of Survival-Associated Genetic Variants under Interval Censoring via the Cox Model

We develop a set of variable selection methods for the Cox model under interval censoring in an ultra-high dimensional setting, where the number of covariates can grow exponentially with the sample size. The proposed methods perform covariate selection via a penalized nonparametric maximum likelihood estimation using popular penalty functions, including lasso, adaptive lasso, SCAD, and MCP. We establish that the penalized estimators with folded concave penalties or the adaptive lasso penalty possess the oracle property. Extensive numerical studies demonstrate that the proposed approaches exhibit satisfactory empirical performance across a variety of scenarios. The practical utility of our methods is further illustrated through an application to genome-wide data with interval-censored survival outcomes.

Chi Hyun Lee, Yonsei University, South Korea

TBD

TBD

IP41 — Stein's Method and Asymptotic Theory

June 14 (Sunday) 14:50–16:30 **Room:** 201 Program

Adrian Röllin, National University of Singapore, Singapore

Interplay of vertex and edge dynamics for dense random graphs

Classical voter processes usually assume that individuals (vertices) change their state based on fixed connections (edges) to their neighbours. In reality, however, social ties or network connections often adapt over time, making the network itself dynamic and creating a two-way interaction between individuals and their links. In this talk, I will introduce a model where each vertex can have one of two opinions, and edges can appear or vanish depending on whether the connected vertices share the same opinion. Despite this added complexity, we find that if we choose the right scaling and let the number of vertices become large, elegant limiting equations emerge. In particular, the fraction of vertices of each opinion follows a Fisher–Wright diffusion, while the network structure evolves through a stochastic flow influenced by the distribution of opinions. We describe this behaviour using a "graphon-valued diffusion", revealing a rich mathematical framework for co-evolving dynamics in dense random graphs.

Lihu Xu, Michigan State University, United States

Quantitative bound for entropic CLT

By extending the Johnson–Barron projection method from one dimension to high dimensions and utilizing a Wang type dimension-free Harnack inequality, we obtain a new quantitative bound for the entropic central limit theorem under the assumption that the Poincaré inequality holds. We compare our results with recent developments to demonstrate the merits of our approach.

Wenkai Xu, University of Warwick, United Kingdom

Stein discrepancies for testing and model assessments

Stein-based discrepancies provide powerful tools to compare distributions and bound their differences. Through integral probability metric, kernelised Stein discrepancies (KSD) are statistically efficient and computationally powerful tools for testing model goodness-of-fit, especially unnormalised models. In (deep) generative models, the model may not come in the form of explicit probability distributions. We describe the KSD based methods for assessing the quality of trained generative models, both for graphs and Euclidean data. In contrast to classical testing, the sample size from the generative models can be as large as desired, while the size of the observed data that the generator aims to emulate is fixed. In this setting classical statistical procedures for assessing model goodness-of-fit may not be applicable. The presentation focus on the non-balanced sample size and conditional score estimation for the development.

Yuta Koike, The University of Tokyo, Japan

High-dimensional third order Edgeworth expansion by Stein's method

Edgeworth expansions refine the central limit theorem by providing higher-order corrections. However, traditional approaches, such as Fourier analytic methods, are not applicable in high-dimensional settings where the dimension tends to infinity as the sample size increases. To avoid this problem, Stein's method has emerged as a powerful alternative approach. In this talk, assuming the existence of Stein kernels, we show the validity of a third order Edgeworth expansion for a sum of independent random vectors in high-dimensional settings. As a byproduct, we obtain an improved convergence rate of Gaussian approximation when the skewness vanishes. This is a joint work with Kouki Akatsuka.

IP35 — Statistical Methods for High-dimensional and Dependent Data

June 14 (Sunday) 14:50–16:30 **Room:** 202 Program

Jun Song, Korea University, South Korea

Sufficient Dimension Reduction via Dependence–Independence Classification

Sufficient dimension reduction (SDR) aims to extract low-dimensional representations of high-dimensional data that retain all information relevant to a response variable. While classical methods are largely restricted to linear structures, modern nonlinear approaches—such as kernel and deep learning methods—often lack a unified formulation and rigorous theoretical guarantees.

In this talk, we present a unified framework for SDR based on a novel perspective that recasts the problem as a binary classification task distinguishing dependent from independent samples. This viewpoint reveals a fundamental connection between SDR and density ratio estimation, showing that the optimal classifier encodes the conditional relationship between predictors and response.

Building on this idea, we propose a neural network–based method that simultaneously learns a low-dimensional representation and a flexible discriminator in an end-to-end manner. We establish theoretical guarantees, including exhaustive recovery of the sufficient structure at the population level and consistency of the estimator under a logistic loss formulation.

Numerical experiments demonstrate strong performance across a range of settings, and a real data analysis illustrates the practical advantages of the proposed approach.

Kyongwon Kim, Yonsei University, South Korea

Learning causal graphs via nonlinear sufficient dimension reduction

We introduce a new nonparametric methodology for estimating a directed acyclic graph (DAG) from observational data. Our method is nonparametric in nature: it does not impose any specific form on the joint distribution of the underlying DAG. Instead, it relies on a linear operator on reproducing kernel Hilbert spaces to evaluate conditional independence. However, a fully nonparametric approach would involve conditioning on a large number of random variables, subjecting it to the curse of dimensionality. To solve this problem, we apply nonlinear sufficient dimension reduction to reduce the number of variables before evaluating the conditional independence. We develop an estimator for the DAG, based on a linear operator that characterizes conditional independence, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated Markov equivalence class. We introduce a modified PC-algorithm to implement the estimating procedure efficiently such that the complexity depends on the sparseness of the underlying true DAG. We demonstrate the effectiveness of our methodology through simulations and a real data analysis.

Junho Yang, Academia Sinica, Taiwan

Denoising spatial point pattern data

In real-life applications, spatial point pattern data are often subject to data manipulations, either unintentionally or by design. In this presentation, we develop a deconvolution method for spatial point processes to extract the features of the true underlying process from its contaminated counterpart. We establish asymptotic theory for both the stationary and nonstationary cases, and the results are further supported by a set of simulations.

Byungwon Kim, Kyungpook National University, South Korea

A Missing Value Imputation Method for High-Dimensional Tabular Data Using DeepInsight and Image Inpainting

While the increasing scale and complexity of modern data is a key characteristic of data science, it has exacerbated the severity of the missing value problem. As the number of features (dimensionality) increases, conventional simple statistical imputation methods distort the complex distribution of the data, while existing methods, such as multivariate imputation algorithms, face challenges of excessive computational cost. To overcome these limitations, this study proposes a novel hybrid approach that applies the image inpainting technique from computer vision to tabular data. The process begins by employing the DeepInsight algorithm to transform the tabular data into image data, where highly correlated features are grouped into the same pixel locations. Following this transformation, a two-stage imputation strategy that leverages the new data structure is performed. For pixels containing a mix of missing and non-missing values, imputation is primarily

performed using information from the non-missing values within that same pixel group. Conversely, when a pixel consists entirely of missing values, it is treated as a 'lost region' of the image, and an AI-based inpainting model is applied to restore its value. Finally, the image with all values imputed is transformed back into tabular data. The proposed method is compared with existing methods using data from The Cancer Genome Atlas (TCGA). By converging technologies from heterogeneous domains, this research presents a new direction for addressing the missing value problem in high-dimensional data.

IP02 — Nonparametric Analysis of Euclidean and Non-Euclidean Data

June 14 (Sunday) 14:50–16:30 **Room:** 214 Program

Andrea Meilan Vila, Universidad Carlos III de Madrid, Spain

Quasi-likelihood estimation for semiparametric circular regression models

Motivated by the need for flexible and interpretable models to handle circular data, this work introduces a semiparametric regression model for a circular response that can include both linear and circular covariates in its parametric and nonparametric components. The nonparametric component allows for modeling complex effects while avoiding restrictive parametric assumptions. Rather than imposing a particular parametric distribution on the error term, we adopt a circular quasi-likelihood formulation, which is useful when the underlying distribution is unknown. Model estimation relies on a backfitting algorithm that iteratively updates the parametric and nonparametric components using circular partial residuals. We establish the asymptotic properties of the resulting parametric and nonparametric estimators and assess their finite-sample performance through simulation studies. An application to the migratory patterns of willow warblers illustrates the advantages of the proposed approach for assessing genetic effects on circular responses and provides new insights into how specific genomic elements influence migratory behaviour.

Gaspard Bernard, Academia Sinica, Taiwan

Testing for sphericity using spatial signs under elliptical directions

It is well known that in a classical elliptical model, testing the rotational symmetry of the underlying distribution is equivalent to testing that a dispersion parameter is a multiple of the identity matrix. We consider the more general model of random vectors with elliptical directions and introduce some scenarios in which testing for rotational symmetry—or at least isotropy—is still equivalent to testing that the dispersion parameter is a multiple of the identity. In particular, since the elliptical direction model allows to depart from the i.i.d. assumption, our approach allows to consider scenarios in which temporal dependence of a certain type is present in the data-generating process. We consider procedures based on the spatial signs of the observations, which can be viewed as a multivariate extension of the classical signs, traditionally used in nonparametric testing for real-valued observations. We argue that, under our assumptions, the classical spatial sign test is a very natural test statistic and we show that, under certain mild conditions, it is asymptotically valid and has the same local asymptotic power as in the classical elliptical scenario. We then show that the spatial sign test is not only robust but also enjoys certain local asymptotic optimality properties when testing for sphericity when the underlying distribution is strongly heavy-tailed.

Giacomo Francisci, University of Trento, Italy

Data depth and dimension reduction

Depth functions are a fundamental tool in nonparametric analysis of multivariate data. They provide a center-outward ordering of observations, which can be used to reveal features of the underlying distribution and as input for further statistical analysis. In particular, depth functions may be employed to estimate multivariate location and to define multivariate quantiles. In some applications, however, it is more natural to consider centrality with respect to a subspace of a given dimension rather than a single point (i.e., a zero-dimensional subspace). We provide a general framework for constructing statistical depth functions that attain their maximum value on a subspace, yielding a center-outward ordering from that

subspace. The directions of the subspace are determined by minimizing a dispersion measure based on the depth function. For elliptically symmetric distributions, these directions coincide with the minimal directions obtained via Principal Component Analysis (PCA). Whereas PCA requires the existence of the covariance matrix, the depth-based approach is fully nonparametric and applicable to distributions of arbitrary shape.

Stanislav Nagy, Charles University, Czech Republic

Computing depth for directional data

Statistical depths introduce notions of nonparametric inference, such as the median, the inter-quantile regions, or ranks, also for data living in multivariate or non-Euclidean spaces. In this talk we focus on directional data, and the two classical depth functions defined in this setup: (i) the angular halfspace (Tukey) depth, and (ii) the angular simplicial depth. Exact computation of both has long been considered infeasible due to prohibitive complexity of standard algorithms. We present a novel geometric approach that yields the first practical exact algorithms for both depths and demonstrate their superior performance over existing methods.

Joint work with Erik Mendroš, Marek Hubař, and Rainer Dyckerhoff.

CS04 — Classification, Variable Selection, and Deep Learning

June 14 (Sunday) 14:50–16:30 **Room:** 215 Program

Bradley Rava, University of Sydney Business School

Ask for More Than Bayes Optimal: A Theory of Indecisions for Classification

Selective classification is a powerful tool for automated decision-making in high-risk scenarios, allowing classifiers to act only when confident and abstain when uncertainty is high. Given a target accuracy, our goal is to minimize indecisions, observations we do not automate. For difficult problems, the target accuracy may be unattainable without abstention. By using indecisions, we can control the misclassification rate to any user-specified level, even below the Bayes optimal error rate, while minimizing overall indecision mass.

We provide a complete characterization of the minimax risk in selective classification, establishing continuity and monotonicity properties that enable optimal indecision selection. We revisit selective inference via the Neyman-Pearson testing framework, where indecision enables control of type 2 error given fixed type 1 error probability. For both classification and testing, we propose a finite-sample calibration method with non-asymptotic guarantees, proving plug-in classifiers remain consistent and that accuracy-based calibration effectively controls indecision mass. In the binary Gaussian mixture model, we uncover the first sharp phase transition in selective inference, showing minimal indecision can yield near-optimal accuracy even under poor class separation. Experiments on Gaussian mixtures and real datasets confirm that small indecision proportions yield substantial accuracy gains, making indecision a principled tool for risk control.

Dongha Kim, Sungshin Women's University

Anomaly Detection by Exploting Training Dynamics in Deep Generative Models

Modern AI systems rely on massive datasets, making them increasingly vulnerable to data quality issues such as noise, missing values, and anomalous observations. This talk introduces an anomaly detection framework based on the inlier-memorization (IM) effect, a training phenomenon in deep generative models where normal patterns are learned earlier than irregular ones. By exploiting this dynamic, we develop efficient strategies that improve detection performance in both unsupervised and semi-supervised settings. The resulting approach enhances robustness, stability, and computational efficiency across diverse data modalities, offering a practical solution for reliable AI deployment.

Hirofumi Ota, University of Tokyo

Fixed-Level Calibration of the Cauchy Combination Test

The Cauchy combination test (CCT) is widely used because it gives a closed-form combined p-value and is known to be asymptotically valid as the nominal level $\alpha \downarrow 0$ under broad dependence structures. We study a different asymptotic question: whether the usual Cauchy cutoff remains accurate at an ordinary fixed level when the number K of combined p-values grows under dependence. Under a canonical one-factor equicorrelated Gaussian copula model, we show that the raw CCT is generally not asymptotically exact at fixed α. With fixed positive correlation, the statistic converges to a random latent-factor limit, so there is no universal fixed-level reference law. When the common correlation ρ_K weakens with K, fixed-level behaviour is governed by the boundary-layer scale $s_K = \sqrt{\rho_K}(\log K)^{3/2}$, and the raw CCT is asymptotically exact if and only if $\rho_K(\log K)^3 \to 0$. Because the size distortion arises entirely from the reference law and not from the statistic, it can be corrected without modifying the test statistic itself. We propose the boundary-layer calibrated CCT (BL-CCT), which replaces the standard Cauchy reference by a one-parameter Gaussian-smoothed Cauchy family while keeping the statistic unchanged. This reference-law correction is fundamentally different from existing approaches that modify the test statistic. BL-CCT is asymptotically exact under the weaker condition $\rho_K \log K \to 0$ and provides a useful finite-K approximation on bounded boundary layers. Numerical experiments support the theory.

Tao He, San Francisco State University

Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data

The adaptive immune system, shaped by somatic V(D)J recombination, may offer prognostic or predictive biomarkers through VJ gene usage. However, analyzing the immune repertoire is challenging due to clonotype heterogeneity. To address this, we propose a novel ensemble feature selection approach and customized statistical learning algorithm focused on VJ gene usage. Applied to TCR sequences from recovered COVID-19 patients, healthy donors, and lung cancer patients receiving immunotherapy, our method identified distinct VJ gene usage patterns linked to recovery and clinical response. Simulation studies show our approach outperforms existing feature selection methods in efficiency, accuracy, stability, and sensitivity, with lower false discovery rates. When combined with classification models, it achieved superior predictive accuracy. Our method effectively classifies immune subtypes, providing insights into immune response signatures to improve treatment strategies.

DL21 — Advancing Spatial Statistical Inference Through Machine Learning, AI, and Modern Computational Methods

June 14 (Sunday) 16:50–18:30 **Room:** LT1A Program

Douglas Nychka, Colorado School of Mines, United States

Mining AI for statistical computation.

One success of statistical science and in particular the analysis of environmental data is the richness of our models to describe spatial dependence including the influence of covariates and the distribution of extreme observations. These more complex models although attractive have come with computational challenges that make them difficult to use. For example, non-stationarity and non-Gaussian fields are common for physical processes and important features for quantifying uncertainty of predictions. However, data analysis with these extensions using traditional statistical computation is time consuming and difficult to implement. The computational technology that has been developed with AI applications in mind can be refactored to advance statistical science. In this talk we suggest a different approach to fitting complex models when one can simulate from the statistical model efficiently. This is often the case, even for models that might not have a closed form likelihood. The idea is that in a Bayesian setting there is a smooth transformation, albeit unknown and possibly complicated, that maps the observed data into the posterior mean of the parameters in the statistical model. One represents this transformation as a neural network (NN) and trains it on a large simulated set of observations. This talk will give two examples of this approach for fitting spatial extremal fields and estimating non-stationary covariance functions. In either case the speedup using NNs is a factor of 100 or more in computation. In this way AI has not replaced spatial statistical analysis but rather contributed new computational tools.

Bo Li, Washington University in St. Louis, United States

Spatially Varying Deep Functional Neural Network: Application in Large-Scale Crop Yield Prediction

Accurate prediction of crop yield is critical for supporting food security, agricultural planning, and economic decision-making. However, yield forecasting remains a significant challenge due to the complex and nonlinear relationships between weather variables and crop production, as well as spatial heterogeneity across agricultural regions. We propose a Spatially Varying Deep Functional Neural Network (SVD-funNet), a deep neural network architecture that integrates functional and scalar predictors with spatially varying coefficients and spatial random effects. The method is designed to flexibly model spatially indexed functional data, such as daily temperature curves, and their relationship to variability in the response, while accounting for spatial correlation.

SVD-funNet mitigates the curse of dimensionality through a low-rank structure inspired by the spatially varying functional index model (SVFIM). Through comprehensive simulations, we demonstrate that SVD-funNet outperforms state-of-the-art functional regression models for spatial data, when the functional predictors exhibit complex structure and their relationship with the response varies spatially in a potentially nonstationary manner.

Application to corn yield data from the U.S. Midwest demonstrates that SVD-funNet achieves superior predictive accuracy compared to both leading machine learning approaches and parametric statistical models. These results highlight the model's robustness and its potential applicability to other weather-sensitive crops.

Yeseul Jeon, Texas A&M University, United States

Uncertainty-Aware Neural Multivariate Geostatistics

We propose Deep Neural Coregionalization, a scalable framework for uncertainty-aware multivariate geostatistics. DNC models multivariate spatial effects through spatially varying latent factors and loadings, assigning deep Gaussian process (DGP) priors to both the factors and the entries of the loading matrix. This joint construction learns shared latent spatial structure together with response-specific, location-dependent mixing weights, enabling flexible nonlinear and space-dependent associations within and across variables. A key contribution is a variational formulation that makes the DGP to deep neural network (DNN) correspondence explicit: maximizing the DGP evidence lower bound (ELBO) is equivalent to training DNNs with weight decay and Monte Carlo (MC) dropout. This yields fast mini-batch stochastic optimization without Markov Chain Monte Carlo (MCMC), while providing principled uncertainty quantification through MC-dropout forward passes as approximate posterior draws, producing calibrated credible surfaces for prediction and spatial effect estimation. Across simulations, DNC is competitive with existing spatial factor models, particularly under strong nonstationarity and complex cross-dependence, while delivering substantial computational gains. In a multivariate environmental case study, DNC captures spatially varying cross-variable interactions, produces interpretable maps of multivariate outcomes, and scales uncertainty quantification to large datasets with orders-of-magnitude reductions in runtime.

IP62 — Frontiers in Statistical Modeling for Complex Data Structures

June 14 (Sunday) 16:50–18:30 **Room:** LT1B Program

Yuhong Yang, Tsinghua University, China

On Mixture of Experts and Local Model Averaging

Mixture of experts (MOE) plays an essential role in LLMs. It is closely related to localized model averaging. In this talk, we will examine benefits of combining models in a way that allows the weights to depend on the input (i.e., localized model averaging), establish oracle inequalities for certain simplified MOE models, and discuss natures of several key features in MOE such as shared experts and top-k in routing.

Mengying You, Shanghai University of International Business and Economics, China

Low-Rank Spatio-Temporal State Space Models with Structured Spatial Covariance

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.

Long Feng, Nankai University, China

High dimensional alpha test for linear factor pricing model with L_q-norm

We consider testing zero pricing errors in high-dimensional linear factor pricing models. Existing methods are mainly based on either an L_2 statistic, which is effective under dense alternatives, or an L_∞ statistic, which is powerful under very sparse alternatives. To bridge these two regimes, we develop a class of L_q-based tests for finite q, including the practically useful L_4 and L_6 cases. We show that larger q leads to greater sensitivity to sparse alternatives. We further establish the asymptotic independence between the L_∞ statistic and the L_q statistic for any finite q, which motivates a Cauchy combination test that adapts to a broad range of sparsity levels. Simulation studies and a real-data analysis show that the proposed methods are more robust to the unknown sparsity of the alternative and can outperform existing procedures in finite samples.

Xu Guo, Beijing Normal University, China

Inference of high-dimensional weak instrumental variable regression models without ridge-regularization

Inference in instrumental variable regressions is challenging in modern applications involving many weak and high-dimensional instruments. A prominent line of work extends the classical Anderson–Rubin test to such settings through ridge regularization. In this paper, we show that ridge regularization is not essential for conducting inference that is robust to weak identification and heteroskedasticity. We propose a tuning-free jackknifed Anderson–Rubin quadratic-form test that is computationally simple and remains feasible even when the number of instruments exceeds the sample size. We further modify the Sup Score test by employing a Gaussian multiplier bootstrap to calibrate critical values under the null. This approach accommodates cross-instrument dependence and typically yields less conservative critical values. The resulting bootstrap-calibrated Sup Score test is particularly powerful when the first-stage relationship between endogenous regressors and instruments is sparse. We establish asymptotic size control for both the tuning-free jackknifed Anderson–Rubin quadratic-form test and the bootstrap-calibrated Sup Score test under arbitrarily weak identification and heteroskedastic errors, and characterize their power properties under alternative hypotheses. Simulation studies and an empirical application demonstrate the practical advantages of the proposed methods relative to existing approaches

IP48 — Markov Chain Simulation

June 14 (Sunday) 16:50–18:30 **Room:** 209A Program

Michael Choi, National University of Singapore, Singapore

Group-averaged Markov chains II: tuning of group action in finite state space

We study group-averaged Markov chains obtained by augmenting a π-stationary transition kernel P with a group action on the state space via orbit kernels. Given a group $\mathcal{G}$ with orbits $(\mathcal{O}_i)_{i=1}^k$, we analyse three canonical orbit kernels: namely the Gibbs (G), Metropolis–Hastings (M), and Barker (B) kernels, as well as their multiplicative sandwiches QPQ and the additive mixtures $\frac{1}{2}(P+Q)$ where $Q \in \{G, M, B\}$. We show that $M^t, B^t \to G$ blockwise as $t \to \infty$ under suitable conditions, that the projection chains induced by $(\mathcal{O}_i)_{i=1}^k$ coincide for GPG and P, and that orbit averaging never deteriorates the absolute spectral gap or asymptotic variance when P is reversible. We give a direct and simple proof of Pythagorean identity under the Kullback-Leibler (KL) divergence, showing that GPG arises naturally as an information projection of P

onto the set of G-invariant transition matrices. For a given P, we characterise the optimal choice of G with a fixed number of orbits that minimises the one-step KL divergence to stationarity. Analogously, for a given G, we characterise the optimal choice of P and give sufficient conditions under which $GPG = \Pi$. We further show that alternating projections over multiple group actions converge at a rate governed by the singular values of an overlap matrix, and that in structured cases, this yields exact sampling where the number of group actions grows logarithmically with the size of the state space. Based on the theory, we propose two heuristics to tune G in practice. We also illustrate the results on discrete uniform and multimodal examples, including the Curie-Weiss model where GPG achieves polynomial (in inverse temperature and dimension) mixing while Glauber dynamics remains exponentially slow.

Cosme Louart, The Chinese University of Hong Kong-Shenzhen, China

Conditions for a Central Limit Theorem for Regularized M-Estimators with General Convex Losses

This talk will present recent results establishing a central limit theorem for projections of regularized M-estimators under general smoothness assumptions on the loss functions and concentration assumptions on the data. The CLT is obtained through leave-one-out techniques by applying a recent result of Shao and Zhang, which simplifies the use of Chatterjee-type Wasserstein bounds.

After outlining the main ingredients of the CLT result, we will explain how to estimate the mean and covariance of the M-estimator, and discuss the consequences for the score and, consequently, for the performance of empirical risk minimization.

Sheng Jiang, The Chinese University of Hong Kong-Shenzhen, China

Errors-in-variables Gaussian Processes for Mixed-input Regression

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.

IP50 — Asymptotic Statistics

June 14 (Sunday) 16:50–18:30 **Room:** 209B Program

Yunyi Zhang, The Chinese University of Hong Kong-Shenzhen, China

Quadratic forms of high-dimensional non-stationary time series: Theory and Applications

Analysis of time series data often relies on estimation and statistical inference for functions of covariance, which in turn requires understanding the behavior of quadratic forms of time series data. This presentation introduces several theoretical results on high-dimensional time series, including concentration inequalities, Gaussian approximation theorem, and variance estimation for quadratic forms. Building on these results, we then develop ANOVA procedures and independence tests for high-dimensional vector time series.

Nan Zou, Macquarie University, Australia

Bootstrap for Dynamical Systems

After its establishment in the late 19th century through the efforts of Poincaré and Lyapunov, the theory of dynamical systems was applied to study processes that change over time. Despite their deterministic nature, dynamical systems can exhibit incomprehensibly chaotic behaviours and seemingly random patterns. Consequently, a dynamical system is usually represented by a probabilistic model, in which the unknown parameters must be estimated using statistical methods. To measure the uncertainty of such estimation, we develop a bootstrap method and establish its consistency and second-order efficiency via continuous Edgeworth expansions. This is a joint work with K. Fernando.

Yuqian Zhang, Renmin University of China, China

Data integration using covariate summaries from external sources

In modern data analysis, information is frequently collected from multiple sources, often leading to challenges such as data heterogeneity and imbalanced sample sizes across datasets. Robust and efficient data integration methods are crucial for improving the generalization and transportability of statistical findings. In this work, we address scenarios where, in addition to having full access to individualized data from a primary source, supplementary covariate information from external sources is also available. While traditional data integration methods typically require individualized covariates from external sources, such requirements can be impractical due to limitations related to accessibility, privacy, storage, and cost. Instead, we propose novel data integration techniques that rely solely on external summary statistics, such as sample means and covariances, to construct robust estimators for the mean outcome under both homogeneous and heterogeneous data settings. Additionally, we extend this framework to causal inference, enabling the estimation of average treatment effects for both generalizability and transportability.

Yaoming Zhen, The Chinese University of Hong Kong-Shenzhen, China

Probabilistic PCA on tensors

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

IP46 — High-dimensional Inference and Dependent Data Analysis

June 14 (Sunday) 16:50–18:30 **Room:** 201 Program

Shinpei Imori, Hiroshima University, Japan

Greedy algorithms in high-dimensional linear regression models with group structure

In the present talk, we consider a variable selection problem in high-dimensional linear regression models, where the number of regression coefficients is allowed to be larger than the sample size. Moreover, we assume that the regression coefficients have group structure, which indicates that the regression coefficients are separated into distinct sub-vectors, and each of sub-vectors is regarded as a component of the model. In order to avoid the computational burden due to the high-dimensionality, we use a greedy algorithm to select the best model. By extending the previous results without considering group structure to our setting, we investigate the convergence rate of our greedy algorithms under sparsity conditions for the grouped regression coefficients.

Yan Liu, Waseda University, Japan

Testing for covariance structures in high-dimensional time series

We consider the testing problem for the sphericity hypothesis regarding the covariance matrix of high-dimensional time series. With the regime of (n, p)-asymptotics, we derive the asymptotic null distributions of U- and V-statistics, which play the main role when the data dimension is large. We propose a spherical bootstrap method for high-dimensional time series for the practical use of these statistics. The numerical simulations align well with our theoretical findings. Some real data applications are also provided.

Hsueh-Han Huang, Academia Sinica, Taiwan

High-dimensional transfer learning using greedy algorithms

We study the problem of high-dimensional transfer learning. Unlike most existing approaches, which primarily rely on Lasso-type methods, we propose a forward-regression-based framework. The sparsity assumptions adopted in this work are more flexible than those typically imposed in the literature, thereby accommodating a wider range of transfer scenarios. Within this framework, we derive sharper convergence rates for the target (gold) estimator under various settings. Our methodology applies to both linear regression and generalized linear models. Numerical experiments further demonstrate the empirical advantages of the proposed approach.

Valentin Patilea, CREST & ENSAI, France

Testing the mean of multivariate random functions

The problem of testing linear hypotheses for the mean functions of random functions is considered. This includes testing whether the mean is zero, whether two sample means are equal, and whether two means differ by a constant shift or ratio. The random functions are defined on a multidimensional compact domain, and multiple independent realizations are observed at random design points, possibly with heteroscedastic errors. The number of design points per realization may be either bounded or arbitrarily large. For two-sample tests, the samples may be unbalanced and dependent. The proposed testing approach is based on a non-asymptotic Gaussian approximation for the estimated Fourier coefficients. Two pivotal chi-square−type test statistics are introduced. The resulting nonparametric tests are powerful and, under conditions relevant to practical applications, can achieve parametric rejection rate. The extension to Hilbert space-valued random functions is also discussed.

IP13 — Innovations in Bayesian Nonparametrics

June 14 (Sunday) 16:50–18:30 **Room:** 202 Program

Li Ma, The University of Chicago, United States

Assessing generative models through density ratio learning

Generative models have become a popular tool for simulating observations from complex, high-dimensional distributions. However, effectively assessing the quality of generative models remains a challenge. Standard approaches, such as reporting empirical distance estimates between real and generated samples, generally do not reveal the nature of the discrepancy, nor do they allow for reliable uncertainty quantification. We show that a powerful approach to understanding such discrepancies is to learn the density ratio between the generative model and the true data-generating model. We demonstrate several strategies for characterizing the discrepancy once reliable density ratio estimates and uncertainty quantification become available. Our starting point is a new loss function for learning density ratios under a two-sample design, based on the variational form of the Hellinger distance. Drawing a connection between this loss and the exponential loss commonly used in training additive tree ensembles for supervised learning, we show that the new loss can be used to train flexible function approximators such as tree ensembles and neural networks. We show that, for additive tree ensembles, tailored

training algorithms that parallel boosting and BART are available, which help attain calibrated generalized Bayesian uncertainty quantification. For deep neural network–based approaches trained using standard gradient descent, we provide some theoretical guarantees in terms of consistency and convergence rates. We demonstrate how to use the inferred density ratios to carry out generative model diagnostics in the context of image data and microbiome compositions.

Igor Prünster, Bocconi University, Italy

Multivariate species sampling models

Species sampling processes have long provided a fundamental framework for random discrete distributions and exchangeable sequences. However, analyzing data from distinct, yet related, sources, requires a broader notion of probabilistic invariance, with partial exchangeability as the natural choice. Over the past two decades, numerous models for partially exchangeable data, known as dependent nonparametric priors, have emerged, including hierarchical, nested, and additive processes. Despite their widespread use in Statistics and Machine Learning, a unifying framework remains elusive, leaving key questions about their learning mechanisms unanswered.

We fill this gap by introducing multivariate species sampling models, a general class of nonparametric priors encompassing most existing dependent nonparametric processes. These models are defined by a partially exchangeable partition probability function, encoding the induced multivariate clustering structure. We establish their core distributional properties and dependence structure, showing that borrowing of information across groups is entirely determined by shared ties. This provides new insights into their learning mechanisms, including a principled explanation for the correlation structure observed in existing models.

Beyond offering a cohesive theoretical foundation, our approach serves as a constructive tool for developing new models and opens new research directions aimed at capturing even richer dependence structures.

Junyi Zhang, The Education University of Hong Kong, Hong Kong

Hierarchical Modelling in Bayesian Factor Analysis

Bayesian factor models are a popular tool for factor analysis. Current state-of-the-art Bayesian factor analysis approaches leverage the beta-Bernoulli process prior to characterize the factors and do not require prior knowledge about the factor dimensionality. This prior, however, ignores the potential hierarchical structure within the factor values, a key aspect for a principled interpretation of the analysis. In this presentation, we introduce a new framework based on a new class of nonparametric priors termed beta-NRMI processes that overcome this limitation. This class of priors allows the development of an innovative hierarchical modelling methodology for Bayesian factor analysis. We present numerical implementations based on simulated and real-world datasets to illustrate the usefulness of our hierarchical modelling methodology.

CS10 — Semiparametric Regression and Censored Data Analysis

June 14 (Sunday) 16:50–18:30 **Room:** 214 Program

Zhi Yang Tho, The Australian National University

Joint Mean and Correlation Regression Models for Multivariate Data

We propose a joint mean and correlation regression model for multivariate discrete and (semi-)continuous response data, that simultaneously regresses the mean of each response against a set of covariates, and the correlations between responses against a set of similarity / distance measures. A set of joint estimating equations are formulated to construct an estimator of both the mean regression coefficients and the correlation regression parameters. Under a general setting where the number of responses can tend to infinity, the joint estimator is demonstrated to be consistent and asymptotically normally distributed, with differing rates of convergence due to the mean regression coefficients being heterogeneous across responses. An iterative estimation procedure is developed to obtain parameter estimates in the required (constrained) parameter space. Simulations demonstrate the strong finite sample performance of the proposed estimator in terms of point estimation and inference. We apply the proposed model to a count dataset of 38 Carabidae ground beetle species sampled throughout Scotland, along with information about the environmental conditions of each site and the traits of each species. Results

show the relationship between mean abundance and environmental covariates differs across the beetle species, and that beetle total length is important in driving the correlations between species.

Sangbum Choi, Korea University

Identifiability and Inference of Semiparametric Copula-Based Quantile Regression under Dependent Censoring

Dependent censoring poses a fundamental non-identifiability challenge in survival analysis, as the joint distribution of failure and censoring times cannot be uniquely determined from observed data. While recent breakthroughs achieve identifiability using parametric or proportional hazards marginals, these models often fail to capture complex covariate effects across survival quantiles. This article proposes a novel semiparametric copula approach where the failure time marginal is modeled through a semiparametric quantile regression framework. We assume a parametric model for the censoring distribution and characterize their dependence using a parametric copula family. We rigorously demonstrate that this model is identifiable without requiring a pre-specified association parameter. To estimate the model components, we develop a computationally efficient method leveraging martingale-based ideas and establish the consistency and asymptotic normality of the resulting estimators. Simulations confirm that our approach provides robust inference for survival quantiles, even when proportional hazards assumptions are violated. The proposed framework offers a more granular and flexible understanding of covariate impacts on survival outcomes under dependent censoring.

Arun Kaushik, Banaras Hindu University, Varanasi, India

Estimation of the Generalized Process Capability Index C_{pyk} under Generalized Progressive Hybrid Censoring using Maximum Likelihood, Robust and Bayesian Approaches

This paper develops a comprehensive inferential framework for the generalized process capability index C_{pyk} under a generalized progressive hybrid censoring (GPHC) scheme. The exponentiated exponential (EE) distribution is adopted as the underlying lifetime model due to its flexibility in modeling skewed industrial data. We derive classical estimation methods, including Maximum Likelihood Estimation (MLE) and a robust alternative. To address the complexities of the GPHC likelihood and ensure reliable interval estimation, a Bayesian framework is developed using a Metropolis-Hastings algorithm. Bootstrap and highest posterior density (HPD) intervals are constructed for interval estimation. Extensive simulation studies demonstrate that the Bayesian approach yields superior performance in terms of coverage probability, while robust estimates outperforms MLE in small, heavily censored samples. The methodology is illustrated using two real datasets from the textile and electronics industries.

Xiaoxi Zhang, Seattle University

A Flexible Test for Comparing Two Hazard Rate Functions with Arbitrary Differences

In medical studies, comparison of two hazard rate functions is fundamentally important when analyzing survival data and evaluating treatment effects. Traditional comparison methods, such as the log-rank, Gehan-Wilcoxon, and Peto-Peto tests, work well for comparing two proportional hazard rates, but could fail when the two hazard rates cross each other or demonstrate a treatment time-lag effect. In practice, however, crossing hazard rates and treatment time-lag effects are fairly common. To study the related research problems with such difference patterns in the hazard rates, some statistical methods have been developed using either parametric survival modeling or nonparametric testing. But, these methods cannot detect both types of difference patterns effectively. In addition, it is usually difficult to justify their parametric model assumptions or derive the null distributions of their nonparametric test statistics. Thus, new methods are needed to compare two hazard rates with an arbitrary difference pattern. In this paper, a new combination test is developed, which integrates the traditional log-rank test, a weighted log-rank test, and a specific version of the Fleming-Harrington test, which is good in comparing two hazard rates with a proportional difference, crossing difference, and difference due to a treatment time-lag effect, respectively. Numerical studies show that it works well in various cases considered.

Charles Zhao, University of North Carolina - Chapel Hill

Generalized Fiducial Method for Topic Modeling

Topic Modeling has received enormous attention in recent years for analyzing text data and discovering unknown or not

easily identifiable topics or recurring subjects. It has been used in various research fields, such as image data, genetics, social science, machine learning, natural language processing, etc. In this article, we propose fiducial latent topic model via a generalized fiducial inference method for topic modeling, which is an innovative approach to provide point or interval estimates for word-topic and topic-document matrices with some established theoretical properties. The simulation studies show that our constructed new methods perform very well or better than some existing methods. A real data analysis for data coming from paper abstracts of some statistical journals is conducted using our approach, which shows the effectiveness of our methods for identifying the contents of research papers, thus offers great potential to be integrated with AI for more effective and accurate results.

Coauthors: Grace Smith and Jan Hannig

CS09 — Bayesian Inference, Monte Carlo, and Statistical Estimation

June 14 (Sunday) 16:50–18:30 **Room:** 215 Program

Pushkar Mohan Kale, National University of Singapore, Singapore

Moment Constrained Cutting Feedback for Modular Bayesian Models

Statistical models are often constructed from multiple linked submodels, each informed by different data sources and domain expertise; these submodels are referred to as modules. Misspecification in any module can distort posterior inference and propagate across the entire model. Cutting feedback methods address this by modifying the joint posterior so unreliable modules do not distort inference in trusted components. The standard formulation finds the Kullback–Leibler closest distribution to the full posterior whose marginal for the shared parameters matches the posterior from the trusted module alone. While principled, this formulation poses a severe computational challenge, as the factored structure of the cut posterior requires a naïve nested MCMC strategy in which a separate inner Markov chain is run for every sample drawn in the outer chain, leading to significant computational cost.

An alternative formulation replaces the full distributional constraint on the shared-parameter marginal with a finite set of low-dimensional moment conditions, specifically matching component-wise means and variances to those obtained under the trusted submodel posterior. The Kullback–Leibler closest distribution to the full posterior satisfying these conditions is called the moment-constrained cut (MCC) posterior. A key property of this distribution is that it admits an exponential tilting representation, in which the full posterior is reweighted by an exponential factor involving the moment constraint functions and a vector of Lagrange multipliers. These multipliers are estimated via stochastic gradient descent using the ADAM optimizer, coupled with MCMC updates of the model parameters. This eliminates the nested chain structure entirely, and the algorithm is theoretically justified with established convergence guarantees.

Numerical studies on a Gaussian location model, an epidemiological study of human papillomavirus prevalence and cervical cancer incidence, and a copula time series model show that the moment-based approach closely approximates standard cut posteriors while being much easier to compute. Across all examples, the imposed moment conditions are satisfied with high accuracy, indicating that the stochastic gradient procedure reliably identifies the correct Lagrange multipliers. Extensions to semi-modular inference, which interpolates between the cut and full posteriors through a degree-of-influence parameter, and to systems with more than two modules, are also outlined.

Ka Lok Lam, University of California, Santa Barbara

Probabilistic representation and Monte Carlo method for nonlinear Dirac equations via telegraph particles

Using branching telegraph particles, a type of piecewise-deterministic Markov processes (PDMP), we develop Feynman–Kac-type probabilistic representations for nonlinear Dirac equations, whose oscillatory complex structure and matrix coupling make such representations less straightforward than in more standard diffusion-based settings. In addition to providing a stochastic interpretation of the nonlinear Dirac dynamics, our construction leads naturally to Monte Carlo approximations of the associated solution map, yielding, to our knowledge, the first probabilistic numerical framework for this class of equations. We also present preliminary observations on the effect of parameter choices, together with numerical experiments examining estimator accuracy, conservation properties, and sensitivity to initial conditions. This is joint work with Jean-Pierre Fouque and Tomoyuki Ichiba (UCSB).

Linus David Fromm, University of Otago

Efficiency of MCMC Samplers for Discrete Inverse Problems

Discrete linear inverse problems occur in many branches of science, including (but not limited to) network tomography, ecology, genetics, and public health. In discrete statistical linear inverse problems, we are interested in a count variable that is not observed directly but rather through a corrupted or aggregated count variable. The aim is to make inferences about the latent variable conditional on the observation. In principle, this requires summing over all possible values of the latent variable, which in most cases is infeasible to enumerate due to the sheer size of the solution set. Instead, we can sample from the fibre −solution set. The idea is to run an MCMC over the solution set with carefully selected moves. Computing this set of moves, known as a Markov basis, is the subject of algebraic statistics. However, even with a Markov basis that enables walks connecting any pair of points on the fibre, the sampler based on these moves may be highly inefficient.

We study the mixing times of samplers based on Markov bases and propose other approaches that experience shorter mixing times. Some of these approaches avoid calculating a Markov basis altogether, while others use different probability weights on elements of a particular basis to improve sampling of the latent variable.

Jyun-Yu Chen, Academia Sinica

Efficient and Interpretable Mixtures of Experts: Statistical Inference, Initialization, and Applications

Scientific discovery increasingly relies on methods that are both flexible and interpretable. Traditional statistical models offer interpretability but depend on restrictive assumptions, whereas modern machine learning methods often sacrifice transparency for predictive accuracy. The Mixture-of-Experts (MoE) framework provides a principled compromise by combining multiple local regression models through a data-driven gating mechanism, capturing complex heterogeneous relationships while remaining interpretable. Despite this potential, practical MoE modeling is hindered by high computational cost, sensitivity to initialization, difficulty in selecting the number of experts, and limited tools for inference and implementation. We address these issues by developing a supervised-compression initializer that provides a data-driven initial guess for the number of clusters. We further propose likelihood-based procedures for confidence intervals and a fixed-gate bootstrap method for prediction intervals. Extensive simulations and case studies show that our MoE framework attains predictive performance comparable to or better than classical nonparametric and machine learning approaches, while providing clearer interpretability and more reliable extrapolation in data-scarce regions. Meanwhile, it achieves over an order-of-magnitude speedup compared with existing packages. These results demonstrate that MoE can serve as a flexible, interpretable, and computationally efficient tool for modern data analysis.

DL07 — To be confirmed

June 15 (Monday) 09:00–10:40 **Room:** LT1A Program

George Michailidis, University of California, Los Angeles, United States

TBD

TBD

Yao Zheng, University of Connecticut, United States

TBD

TBD

DL08 — To be confirmed

June 15 (Monday) 09:00–10:40 **Room:** LT1B Program

Xiao Wang, Purdue University, United States

Coreset-Induced Flow Matching

Flow matching has emerged as a flexible framework for training continuous generative models, but its statistical and computational behavior depends critically on the choice of coupling between source and target samples. In this talk, we introduce coreset-induced flow matching, a framework that constructs couplings through a weighted coreset representation of the target distribution together with assignment-based local bridge distributions. We study the resulting method from both approximation and estimation perspectives, showing how coreset quality and bridge construction affect the learned velocity field and the final generative model. The proposed algorithm combines Sinkhorn coreset selection with local Gaussian bridge sampling and can be implemented within standard flow-matching pipelines. Numerical experiments on synthetic manifolds and image benchmarks illustrate the tradeoff between statistical fidelity and computational efficiency, and highlight when coreset-induced couplings can provide practical advantages over more direct matching schemes.

Yongdai Kim, Seoul National University, South Korea

Uncertainty-adaptive Feedback Guidance for Improved Image Generation with Diffusion Models

The success of modern diffusion models heavily depends on the guidance mechanisms that align generated images with conditions. However, the widely adopted classifier-free guidance (CFG) applies a global guidance scale across the entire image. Hence, it can overlook the complex dynamics of the generative process, and may result in misaligned features or visual artifacts. To overcome this limitation, we introduce a novel guidance mechanism named Uncertainty-adaptive Feedback Guidance (UFG) that leverages pixel-wise uncertainty as an adaptive scale. By utilizing uncertainty estimates, our method adaptively determines the pixel-wise guidance scale by minimizing the one-step marginal variance at each denoising step. This guidance affects the next-step uncertainty estimation, and hence establishes a feedback loop. Through empirical evaluations across various image generation tasks, we demonstrate that scaling guidance based on pixel-wise uncertainty enhances overall generation quality and effectively mitigates visual artifacts. Consequently, UFG achieves consistently strong performance compared to standard CFG and other existing uncertainty-aware baselines.

Yixuan Qiu, Shanghai University of Finance and Economics, China

GPU-Accelerated Solver for Entropic-Regularized Optimal Transport

Optimal transport (OT) has emerged as a fundamental tool in modern machine learning, yet its computational cost remains a significant bottleneck for large-scale applications. While harnessing the massive parallelism of modern GPU hardware is critical for efficiency, the de facto standard Sinkhorn algorithm, despite its ease of parallelization, often suffers from slow convergence in challenging problems. More recently, the sparse-plus-low-rank quasi-Newton method offers a balance between convergence rate and per-iteration complexity; however, its efficiency on GPUs is severely hindered by the serial nature of sparse matrix symbolic analysis and irregular memory access patterns. To bridge this gap, we present cuRegOT, a high-performance GPU solver tailored for entropic-regularized OT. We introduce a suite of algorithmic and architectural optimizations, including an amortized symbolic analysis strategy to mitigate CPU bottlenecks, an asynchronous Sinkhorn iterates generation mechanism, and a fused kernel for bandwidth-efficient gradient evaluation. These strategies are backed by rigorous theoretical guarantees ensuring algorithmic convergence. Extensive numerical experiments demonstrate that cuRegOT achieves significant speedups over state-of-the-art GPU-based solvers across a variety of benchmark tasks.

IP56 — Modern Nonparametric Approaches for Dependent Data Analysis

June 15 (Monday) 09:00–10:40 **Room:** 209A Program

Chae Young Lim, Seoul National University, South Korea

Exploring Spatial Dynamics in Regression Coefficients: A Bayesian Regularization Method with Clustering

Analyzing datasets with spatial information in regression setting often challenges the assumption that the relationship between the response variable and explanatory variables remains homogeneous across the spatial domain. To relax such assumption, we propose a Bayesian Regularized Spatially Clustered Coefficient model, which detects cluster-wise varying

effects and performs variable selection simultaneously. The proposed model identifies key covariates influencing the response variable by introducing a selection prior and uncovers their spatially clustered coefficient effects using a clustering prior. Bayesian inference is implemented via the Reversible Jump Markov chain Monte Carlo (RJMCMC) method, utilizing two types of reversible jump moves that enable efficient exploration of the parameter space. The collapsed posterior and parallel tampering are also considered for better mixing of the RJMCMC samples. A simulation study was conducted under various settings to demonstrate the effectiveness of the proposed approach. Finally, our model was applied to real data from the 19th and 20th presidential elections in South Korea to identify factors influencing the vote share of the elected president and to examine their spatial cluster-specific effects, as the data exhibit strong spatial dependence driven by the pronounced regional characteristics of political preferences.

Arindam Chatterjee, Indian Statistical Institute, Delhi, India

Predicting network summary statistics through network sampling: some rigorous results under induced and egocentric network formation

Consider a large population network which is generated from a sparse Stochastic Block Model. A Bernoulli node sampling scheme is used to select nodes. To quantify the effects of sampling under resource constraints, we allow 'sparse' node sampling, where the node selection probability is allowed to decay to zero, as the network size increases. We study the effects of both induced and egocentric network formation, following the initial sampling of nodes.

For any target subgraph H, we provide thresholds on the model and sampling sparsity levels, which will ensure a Gaussian limit law for the network sample based estimated subgraph count statistic. We also derive quantitative bounds on the speed of Gaussian approximation. It is shown that these thresholds are intricately dependent on the edge density and the minimum vertex cover size of H. We find that if the model sparsity level remains below an initial threshold, the speed of Gaussian approximation is unaffected by the model sparsity level. Beyond this initial threshold, there is a rapid deterioration. As the sparsity levels increase further, in case of strictly balanced H, we obtain various types of Poisson distribution based limit laws. The sufficient conditions for a Gaussian limit law also turn out to be necessary. Using these limit laws we obtain we obtain prediction intervals for the unknown population subgraph count. However, in the egocentric case, there are certain choices of H which are not covered by our results, and one has to use case-by-case arguments. A simulation study strongly supports our theoretical results.

(This is a joint work my PhD student, Mr. Anirban Mandal)

Soudeep Deb, Indian Institute of Management Bangalore, India

Nonparametric regression of spatio-temporal data using infinite-dimensional covariates

In spatio-temporal analysis, we often record data at specific time intervals but with varying spatial locations between these timepoints. We propose a conditional model to analyze such spatio-temporal data that accommodates the dependencies alongside second-order stationary explanatory variables, which may be infinite-dimensional and accommodate spatio-temporal covariates. Because of the absence of a mixing-type dependence condition in this case, which is typically required by the existing studies, we consider a weaker polynomially decaying moment contraction (PMC) condition on the covariates. In this paper, we obtain nonparametric point estimates of the mean and covariate functions of such a regression model, which we then show to be statistically consistent. We also obtain a simultaneous confidence interval of the mean function using the central limit theorem for the proposed estimator. Such simultaneous inference tools can be used to test for certain specifications of the mean function. Some simulation studies and two real-data analyses have been illustrated to corroborate the findings.

Soutir Bandyopadhyay, Colorado School of Mines, United States

Frequency Domain Resampling for Gridded Spatial Data

In frequency domain analysis for spatial data, spectral averages based on the periodogram often play an important role in understanding spatial covariance structure, but also have complicated sampling distributions due to complex variances from aggregated periodograms. In order to nonparametrically approximate these sampling distributions for purposes of inference, resampling can be useful, but previous developments in spatial bootstrap have faced challenges in the scope of their validity, specifically due to issues in capturing the complex variances of spatial spectral averages. As a consequence, existing frequency domain bootstraps for spatial data are highly restricted in application to only special processes (e.g. Gaussian) or certain spatial statistics. To address this limitation and to approximate a wide range of spatial spectral averages,

we propose a practical hybrid-resampling approach that combines two different resampling techniques in the forms of spatial subsampling and spatial bootstrap. Subsampling helps to capture the variance of spectral averages while bootstrap captures the distributional shape. The hybrid resampling procedure can then accurately quantify uncertainty in spectral inference under mild spatial assumptions. Moreover, compared to the more studied time series setting, this work fills a gap in the theory of subsampling/bootstrap for spatial data regarding spectral average statistics.

IP15 — Statistical Machine Learning for High-dimensional Neuroimaging Time Series

June 15 (Monday) 09:00–10:40 **Room:** 209B Program

Michele Guindani, University of California, Los Angeles, United States

Decoding Neuronal Ensembles from Spatially-Referenced Calcium Traces: A Bayesian Semiparametric Approach

Understanding how neurons coordinate their activity is a fundamental question in neuroscience, with implications for learning, memory, and neurological disorders. Calcium imaging has emerged as a powerful method to observe large-scale neuronal activity in freely moving animals, providing time-resolved recordings of hundreds of neurons. However, fluorescence signals are noisy and only indirectly reflect underlying spikes of neuronal activity, complicating the extraction of reliable patterns of neuronal coordination. We introduce a fully Bayesian, semiparametric model that jointly infers spiking activity and identifies functionally coherent neuronal ensembles from calcium traces. Our approach models each neuron's spiking probability through a latent Gaussian process and encourages anatomically coherent clustering using a location-dependent stick-breaking prior. A spike-and-slab Dirichlet process captures heterogeneity in spike amplitudes while filtering out negligible events. We consider calcium imaging data from the hippocampal CA1 region of a mouse as it navigates a circular arena, a setting critical for understanding spatial memory and neuronal representation of environments. Our model uncovers spatially structured co-activation patterns among neurons and can be employed to reveal how ensemble structures vary with the animal's position.

Eardi Lila, University of Washington, United States

Biophysics-informed deep operator learning for electrophysiological source reconstruction

Electrophysiological brain signals are typically acquired through indirect and noisy measurements, providing transformed representations of the underlying neural activity. Source reconstruction — the inverse problem of resolving underlying neural signals from these measurements — is essential for accurate brain function mapping and for advancing our understanding of brain development and disorders. However, this task remains challenging due to its mathematical ill-posedness and resulting sensitivity to noise. These challenges are not unique to electrophysiological source reconstruction but are shared across a broad class of inverse problems, including computed tomography, seismic imaging, and remote sensing. Deep learning methods have shown promise across a range of inverse problems, but they often disregard the underlying biophysical principles governing the data generation process, which can reduce data efficiency. Here, we introduce a biophysics-informed geometric deep operator learning framework (DeepOp-Informed) that embeds biophysical constraints into the model via a custom layer, resulting in more efficient learning and improved reconstruction performance. This custom layer enables the neural network to adapt to subject-specific variations in the physics of signal generation, resulting from differences in brain anatomy, sensor positioning, and the presence of malfunctioning sensors. The proposed method is demonstrated for magnetoencephalography source reconstruction using adolescent data, which poses significant challenges due to low signal-to-noise ratios. While our application focuses on magnetoencephalography, the framework is general and readily extendable to other imaging modalities.

IP58 — Frontiers in Bayesian Learning and Inference

June 15 (Monday) 09:00–10:40 **Room:** 203 Program

Yingzhen Li, Imperial College London, United Kingdom

LLMs as implicit/predictive Bayesian models: algorithmic frontiers

As large language models (LLMs) gain popularity in conducting prediction tasks in-context, handling uncertainty in in-context learning becomes essential to ensuring reliability. The recent hypothesis of in-context learning performing predictive Bayesian inference opens the avenue for Bayesian modelling ideas to be applied to LLMs. However, unlike conventional Bayesian methods with explicit model constructions (parameter specification, prior and likelihood functions), the underlying Bayesian model of an LLM (if it exists due to de Finetti) is implicitly defined, rendering many Bayesian algorithms inapplicable in this context. In this talk, I will discuss two research works from my team on extending algorithms for explicit Bayesian modelling to the implicit Bayesian modelling regime, showing the potential of Bayesian uncertainty quantification and belief updates in auto-regressive LLM settings. The first work introduces a variational uncertainty decomposition framework for in-context learning without explicitly sampling from the latent parameter posterior, by optimising auxiliary queries as probes to obtain an upper bound to the aleatoric uncertainty of an LLM's in-context learning procedure, which also induces a lower bound to the epistemic uncertainty. The second work introduces predictive versions of Thompson sampling for implicit Bayesian models in bandit learning context, which achieves the same expected regret as conventional Thompson sampling with explicit Bayesian models (optimal in no model mismatch setting).

David Frazier, Monash University, Australia

Predictive Bayesian Inference on Population Functionals

Predictive Bayesian inference (PBI) represents a model-and-prior agnostic approach to standard Bayesian inference which allows users to quantify uncertainty for a functional of interest only by specifying a forward predictive model for future unobserved data. The flexibility and generality of this framework have led to a host of novel algorithms for implementing this approach, and many empirical applications, yet the reliability of the resulting inferences for the underlying statistical functional of interest remains unclear. Herein, we demonstrate that when using PBI for a population functional of interest, the resulting posterior concentrates onto a well-defined quantity that explicitly depends on the predictive engine used to implement the predictive recursion underlying the method. Furthermore, this predictive engine entirely determines the uncertainty quantification produced in PBI. Consequently, our results show that if the predictive engine does not capture all relevant features of the data, and ⁻even in very simple examples ⁻the coverage of predictive Bayes credible sets for the population value of the functional of interest can be arbitrarily close to zero. We carefully explain why this occurs, and show that this behavior is directly tied to the inaccuracy of the predictive engine used to produce future observations within the PBI framework. As a consequence, our results imply that in order for PBI to deliver calibrated posterior inferences, the resulting predictive engine used to generate posterior samples must contain, in a well-defined sense, the true DGP, else inferences generated under this framework will not be calibrated.

Susan Wei, Monash University, Australia

Pretrained Transformers as In-Context Bayesian Learners: Implications for Uncertainty Quantification

Foundation models for tabular prediction, such as TabPFN, achieve state-of-the-art accuracy by amortizing Bayesian learning into a single forward pass. But how do we get trustworthy uncertainty estimates for and out of them? In this talk, I show that certain pretrained transformers lend themselves to being interpreted as in-context Bayesian learners ⁻⁻and that this interpretation opens the door to principled uncertainty quantification. First, I introduce TabMGP, which uses TabPFN as a predictive rule within the martingale posterior framework, sidestepping the need for prior and likelihood specification altogether while producing credible sets with near-nominal coverage. Second, I address the problem of decomposing TabPFN's predictive uncertainty into epistemic and aleatoric components ⁻⁻something not previously possible ⁻⁻by casting this as a Bayesian predictive inference problem and proving a predictive CLT under quasi-martingale conditions. The resulting variance estimators are fast, target epistemic uncertainty directly, and again achieve good frequentist coverage.

IP34 — Recent Advances in Stochastic Processes

June 15 (Monday) 09:00⁻10:40 **Room:** 201 Program

Junichiro Yoshida, The University of Tokyo, Japan

Estimation Error and Hypothesis Testing for Non-identifiable Models, with Applications to Machine Learning

To bridge machine learning and classical statistical methods, it is important to analyze estimation error in non-identifiable models such as neural networks. However, in complex non-identifiable models, estimation error is difficult to evaluate explicitly. To address this problem, we propose an estimator of the estimation error together with its confidence interval, thereby enabling classical statistical methods, such as hypothesis testing, for non-identifiable models. The key idea is to apply the resolution of singularities to singular models following Watanabe. While Watanabe's theory is primarily developed in a Bayesian framework, our estimator is feasible within a frequentist asymptotic framework and is computationally tractable.

Jie Yen Fan, Monash University, Australia

Estimation in age-and-population-dependent models

We consider general age-and-population-dependent population systems, where individual birth and death rates depend not only on age but also on the overall population composition. These systems can be modelled as measure-valued stochastic processes. Functional Law of Large Numbers and Central Limit Theorem with large carrying capacity can be established. These results, with appropriate test functions, allow us to estimate the demographic rates and obtain some inference results. Joint work with Kais Hamza, Fima Klebaner and Ziwen Zhong.

Tetsuya Takabatake, The University of Osaka, Japan

Optimal Estimation for General Gaussian Processes in the Frequency Domain

We study frequency-domain estimation for general Gaussian processes when the spectral density is available only through an approximation. We propose an approximate Whittle maximum likelihood method and give general conditions under which it is asymptotically equivalent to the exact likelihood. Under these conditions, the resulting estimator is shown to be consistent, asymptotically normal, and efficient. We further introduce a unified spectral approximation framework that covers a broad class of models and yields a transparent route to verification. The theory is particularly useful for continuous-time processes, where spectral aliasing often makes direct estimation impractical and where truncation-based methods may suffer from substantial finite-sample bias. Numerical experiments demonstrate that the proposed approach effectively mitigates aliasing-induced bias.

Yasutaka Shimizu, Waseda University, Japan

Joint estimation for mean functions and covariance kernels in Gaussian processes

We propose a contrast-based estimation method for Gaussian processes with time-inhomogeneous drifts, observed under high-frequency sampling. The process is modeled as the sum of a deterministic drift function and a stationary Gaussian component with a parametric kernel. Our method constructs a local contrast function from adjacent increments, which avoids inversion of large covariance matrices and allows for efficient computation. We prove consistency and asymptotic normality of the resulting estimators under general ergodicity conditions. A distinctive feature of our approach is that the drift estimator attains a nonstandard convergence rate, stemming from the direct Riemann integrability of the drift density. This highlights a fundamental difference from standard estimation regimes. Furthermore, when the local contrast fails to identify all parameters in the covariance kernel, moment-based corrections can be incorporated to recover identifiability. The proposed framework is simple, flexible, and particularly well-suited for high-frequency inference with time-inhomogeneous structure.

IP54 — Sampling and Optimization in Modern Data Science Problems

June 15 (Monday) 09:00–10:40 **Room:** 202 Program

Ning (Patricia) Ning, Texas A&M University, United States

Robust Iterative Learning Hidden Quantum Markov Models

Hidden Quantum Markov Models (HQMMs) extend classical Hidden Markov Models to the quantum domain, offering a powerful probabilistic framework for modeling sequential data with quantum coherence. However, existing HQMM learning algorithms are highly sensitive to data corruption and lack mechanisms to ensure robustness under adversarial perturbations. In this work, we introduce the Adversarially Corrupted HQMM (AC-HQMM), which formalizes robustness analysis by allowing a controlled fraction of observation sequences to be adversarially corrupted. To learn AC-HQMMs, we propose the Robust Iterative Learning Algorithm (RILA), a derivative-free method that integrates a Remove Corrupted Rows by Entropy Filtering (RCR-EF) module with an iterative stochastic resampling procedure for physically valid Kraus operator updates. RILA incorporates L1-penalized likelihood objectives to enhance stability, resist overfitting, and remain effective under non-differentiable conditions. Across multiple HQMM and HMM benchmarks, RILA demonstrates superior convergence stability, corruption resilience, and preservation of physical validity compared to existing algorithms, establishing a principled and efficient approach for robust quantum sequential learning.

Yuchen Wu, Cornell University, United States

On the Robustness of Distribution Support under Diffusion Guidance

Diffusion guidance is a powerful technique that enables controllable and high-fidelity sample generation with diffusion models. At a high level, it modifies the score function by incorporating a guidance term that steers the generative process toward a desired condition. Despite its empirical success, the theoretical properties of diffusion guidance remain largely unexplored, and it is not well understood why it consistently produces high-quality samples.

In this work, we explain the effectiveness of diffusion guidance by establishing a robustness of support property. Specifically, we show that, given exact access to the score functions, guided diffusion processes almost always generate samples that remain close to the target support. This property is particularly desirable, as samples that lie off the support are often structurally implausible and may adversely affect downstream tasks. Our analysis covers both Denoising Diffusion Implicit Models (DDIM) and Denoising Diffusion Probabilistic Models (DDPM), and applies to a wide range of discretization schemes induced by exponential integrators. Our results provide a rigorous foundation for understanding why diffusion guidance produces physically meaningful and structurally plausible samples.

Rong Tang, The Hong Kong University of Science and Technology, Hong Kong

Robust Bayesian Inference on Riemannian Submanifold

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 much larger ambient dimension. Our numerical results demonstrate the effectiveness of our approach on a variety of problems.

Runmin Wang, Texas A&M University, United States

Change-point detection in high-dimensional time series using MOSUM

In this talk we study the problem of detecting abrupt dense changes in the mean of a high-dimensional time series. We shall focus on the dense change in the sense that a large proportion of the elements in the mean vectors can change, although our

method can also handle sparse change if the jump size is large. Specifically we developed a nonparametric methodology to identify the change-point location for a time series with both temporal and cross-sectional dependence. We construct a MOSUM statistic which can also be considered as a trimmed U-statistic for the l_2 norm of the mean change at each location, and the local maximizer of the MOSUM statistics serves as a natural estimator for the true change point location. We further compare the constructed MOSOM statistics with some threshold to determine the number of changes points in the data so that the same method can work under both single change point model or multiple change point model. Simulation results show that the method can accurately estimate both the number and the location of change points, and the method is not sensitive to the tuning parameters.

IP55 — Modern Perspectives on Causal Inference

June 15 (Monday) 09:00–10:40 **Room:** 214 Program

Raaz Dwivedi, Cornell University, United States

TBD

TBD

Shuangning Li, The University of Chicago, United States

Covariate Adjustment Cannot Hurt: Treatment Effect Estimation under Interference with Low-Order Outcome Interactions

In randomized experiments, covariates are often used to reduce variance and improve the precision of treatment effect estimates. However, in many real world settings, interference between units, where one unit's treatment affects another's outcome, complicates causal inference. This raises a key question: how can covariates be effectively used in the presence of interference? Addressing this challenge is nontrivial, as direct covariate adjustment, such as through regression, can sometimes increase variance due to dependencies across units. In this paper, we study how to use covariate information to reduce the variance of treatment effect estimators under interference. We focus on the total treatment effect (TTE), defined as the difference in average outcomes when all units are treated versus when all are controlled. Our analysis is conducted under the neighborhood interference model and a low order interaction outcome model. Building on the SNIPE estimator from Cortez-Rodriguez et al. (2023), we propose a covariate adjusted SNIPE estimator and show that, under sparsity conditions on the interference network, the proposed estimator is asymptotically unbiased and has asymptotic variance no greater than that of the original SNIPE estimator. This parallels the classical result of Lin (2013) under the no interference assumption, where covariate adjustment does not worsen estimation precision. Importantly, our variance improvement result does not rely on strong assumptions about the covariates: the covariates may be arbitrarily dependent, affect outcomes across units, and depend on the interference network itself.

Rajarshi Mukherjee, Harvard University, United States

Inference in high-dimensional linear mediation models under proportional asymptotics

In this talk, we will consider mediation analysis in a linear structural equation model under a proportional asymptotic regime where both the number of mediators and confounders are allowed to diverge proportionally to the sample size. In this regime, we provide consistent and asymptotically normal estimators of direct, indirect, and suitable path-specific effects through a carefully devised scheme without imposing any sparsity structure on the problem. Indeed, it turns out that typical assumptions made in debiasing methodology under proportional asymptotic regimes do not apply to the mediation analysis problem, and we appeal to a partial ridge-based debiasing scheme to address the relevant subtleties. We additionally consider the effect of sample splitting and cross-fitting and perform detailed numerical experiments to validate our results in finite samples.

Debraj Das, Indian Institute of Technology Bombay, India

Asymptotic Theory of K-fold Cross-validation in Lasso and the Validity of Bootstrap

Least absolute shrinkage and selection operator or the Lasso is one of the widely used regularization methods in regression. Statisticians usually implement Lasso in practice by choosing the penalty parameter in a data-dependent way, the most popular being the K-fold cross-validation (or K-fold CV). However, inferential properties, such as the variable selection consistency and $n^{1/2}$-consistency, of the K-fold CV based Lasso estimator are still unknown. In this paper, we consider the heteroscedastic linear regression model and show that the Lasso estimator with K-fold CV based penalty is $n^{1/2}$-consistent, but not variable selection consistent. Based on $n^{1/2}$-consistency, we also establish the validity of Bootstrap in approximating the distribution of the Lasso estimator. Therefore, our results theoretically justify the use of K-fold CV in choosing the penalty parameter in Lasso and prescribes a way of drawing valid inferences in linear regression based on the CV-based Lasso estimator. Simulation results also justify our theoretical findings in finite samples.

CS03 — High-Dimensional Inference, Factor Models, and Latent Variables

June 15 (Monday) 09:00–10:40 **Room:** 215 Program

Yuqi Zhang, University of Bristol

Multiscale Detection of Multiple Change Points in High-Dimensional Factor Models

We introduce a multiscale, bandwidth-free procedure for detecting multiple change points in large approximate factor models, referred to as FMseg.

We use an observationally equivalent representation of the factor model with structural changes, under which a single transformation matrix keeps the loading matrix globally invariant, so that only the pseudo factor covariances vary across regimes, eliminating repeated eigendecompositions and reducing numerical error. Allowing for serial dependence and possibly heavy-tailed innovations, we show that FMseg consistently recovers both the number and the locations of change points. Simulations demonstrate high detection accuracy and localisation properties of FMseg even when changes are closely spaced in dependent, high-dimensional data.

An empirical analysis of the daily S&P 500 dataset uncovers structural breaks in volatility aligned with major market events.

The R code implementing the proposed method is available at `https://github.com/YuqiZhangSA/FMSeg_vec`.

Kyoowon Kim, Seoul National University

Testing and Segmentation of Joint and Individual Components in Integrative Multi-Source Factor Models

Multi-source data arise when multiple types of measurements are collected from the same set of samples, and they often contain both shared variation across sources and source-specific variation. Identifying the joint structure that is common across sources while separating individual structures specific to each source is a fundamental task in multi-source data integration. Deterministic joint–individual structure models have been proposed to address these issues, but many existing methods rely on computationally intensive optimization procedures or thresholds based on loose spectral bounds, which can lead to reduced detection power and slow computation. In this paper, we propose MSSAT (Multi-Source Sequential Alignment Test), a new method for estimating joint structure in multi-source data. The key idea is to measure the geometric alignment between score subspaces across different data sources. When multiple sources share a joint component, their corresponding subspace directions tend to be closely aligned, whereas directions associated with individual structure exhibit weaker alignment. Building on this observation, we develop a statistical alignment test that determines whether structural components from different sources correspond to a common joint direction. We derive the asymptotic distribution of the proposed alignment statistic and construct a sequential testing procedure for estimating the joint rank. Simulation studies demonstrate that MSSAT achieves stronger detection ability and substantially faster computation compared with existing methods. Applications to several real datasets, including the TCGA multi-omics dataset, demonstrate that the proposed method can be effectively applied to real multi-source data.

Zhining Wang, The Australian National University

Simultaneous Inference for Latent Variable Predictions in Factor Analytic Models

Factor analytic models, also known as latent variable models, are fundamental tools in multivariate statistics and are widely used in fields such as psychology, economics, and the social sciences. Although substantial research has focused on estimation and inference for model parameters such as the loading matrix and error variances, much less attention has been paid to inference for the predicted latent factors themselves. In particular, the problem of constructing joint prediction intervals for latent variables simultaneously across more than one cluster remains underexplored. In this talk, we develop a framework for simultaneous inference on predicted latent factors in factor analytic models. We present methods for constructing simultaneous prediction intervals, with particular attention to bootstrap-based procedures. We also discuss choices of prediction mean squared error and compare the proposed the bootstrap strategy with other "simpler" approaches such as the Bonferroni correction and Monte-Carlo simulation. The practical feasibility and robustness of these methods are illustrated through simulation studies and an application in the biosciences.

Giheon Seong, Seoul National University

James–Stein Estimation of Spiked Eigenvectors Under the Generalized Spiked Population Model

In high-dimensional Principal Component Analysis (PCA), standard sample eigenvectors are notoriously inconsistent estimators of their population counterparts due to the high-dimensional noise. To address this challenge, we propose a James–Stein-type shrinkage estimator for spiked eigenvectors under the generalized spiked population model. Our approach incorporates prior structural information formulated as a deterministic, low-dimensional target subspace. By shrinking the standard sample eigenvectors towards an effective target subspace—spanned by this prior information and the remaining spiked sample eigenvectors—we construct an estimator that achieves a desired property. We theoretically prove that this James–Stein estimator strictly dominates standard PCA when the target subspace is informative, while remaining asymptotically harmless by reducing to standard PCA when the prior is completely uninformative. Furthermore, we establish the robustness of our proposed method by demonstrating that these theoretical guarantees hold across diverse high-dimensional frameworks, including the classical Random Matrix Theory (RMT) regime where the dimension and sample size grow proportionally, the ultra-high-dimensional (UHD) regime, and the high-dimension, low-sample-size (HDLSS) regime.

Kanta Naito, Graduate School of Information Sciences

Aspects of High-Dimensional Kernel Density Estimation: Bandwidth-Induced Bifurcations and Their Estimation

Seventy years have passed since Kernel Density Estimation (KDE) was first proposed in the seminal paper by Rosenblatt (1956). In an era dominated by "big data," does KDE truly function as an effective tool for extracting meaningful information? Assuming that big data is inherently high-dimensional, this study discusses the behavior of the kernel density estimator within a high-dimensional framework. Conventional results for KDE, such as consistency and asymptotic normality, have typically been derived under an asymptotic framework where the bandwidth decreases as the sample size increases while keeping the dimension fixed. However, if this constraint is removed and we move to a framework where the dimension increases alongside the sample size, simultaneously reducing the bandwidth leads to a direct degradation in estimation accuracy. Therefore, a setting where the bandwidth is fixed while both the sample size and dimension increase serves as a more appropriate foundation for investigation. In this case, the specific value of the bandwidth determines a bifurcation in the estimator's behavior in high dimensions. Within this framework, recent studies in statistical physics have investigated the behavior of high-dimensional kernel density estimators. These works employ advanced concepts and methods such as free energy, the replica method, replica free entropy, glass transitions, and large deviation principles to provide deep theoretical insights into KDE behavior. In this talk, building upon these perspectives from statistical physics, we explain how the asymptotic behavior of the kernel density estimator varies depending on the size of the fixed bandwidth. We introduce rigorous results focusing on the bandwidth scales required for the law of large numbers, the central limit theorem, and uniform integrability to hold in high-dimensional settings. Additionally, we discuss the limit expressions of performance metrics such as Mean Integrated Squared Error. We also present an estimator for the solution of the equation governing the bifurcation of the estimator's asymptotic behavior. Results from Monte Carlo simulations and practical examples using real-world data are also reported.

DL17 — Change Point Detection and Its Modern Applications

June 15 (Monday) 11:00–12:40 **Room:** LT1A Program

Jialiang Li, National University of Singapore, Singapore

Change-point Detection and Its Modern Applications

We review recent advances in change-point detection methods across three important fields of statistics: (i) We first present a subgroup identification method based on a multi-threshold change plane model where the subgroup boundaries are defined by a high-dimensional hyperplane in the covariate space. Subjects grouped into different regions may receive more individualized treatments in medical research studies and achieve improved health outcomes. (ii) We then consider the estimation of discontinuity for functional process data. Many longitudinal or functional responses may exhibit abrupt jumps and our methodology effectively accommodates such complicated non-smooth features. (iii) We finally explore change-point estimation within dynamic networks using a recently proposed network autoregressive model. This framework demonstrates that community structures in networks can shift similarly to changes observed in time series data. These reviews highlight the wide-ranging applications of change-point detection methodologies in modern data analysis.

Alex Luedtke, University of Washington, United States

Simplifying debiased inference via automatic differentiation and probabilistic programming

We introduce an algorithm that simplifies the construction of efficient estimators, making them accessible to a broader audience. 'Dimple' takes as input computer code representing a parameter of interest and outputs an efficient estimator. Unlike standard approaches, it does not require users to derive a functional derivative known as the efficient influence function. Dimple avoids this task by applying automatic differentiation to the statistical functional of interest. Doing so requires expressing this functional as a composition of primitives satisfying a novel differentiability condition. Dimple also uses this composition to determine the nuisances it must estimate. In software, primitives can be implemented independently of one another and reused across different estimation problems. We provide a proof-of-concept Python implementation and showcase through examples how it allows users to go from parameter specification to efficient estimation with just a few lines of code.

Jessica Li, University of California, Los Angeles, United States

TwinPoSI: A Synthetic Data-Based Method for Valid and Powerful Post-Selection Inference

Post-selection inference aims to provide valid uncertainty quantification after data-driven feature selection, but is often challenged by selection-induced bias and complex post-selection distributions. In this talk, I will introduce TwinPoSI, a novel framework for post-selection inference based on the concept of statistical twins. The key idea is to construct synthetic twin datasets that mimic the distributional structure of the original data under a specified model, enabling valid confidence intervals for features selected by LASSO. TwinPoSI achieves valid control of the False Coverage Rate (FCR) for population-based regression coefficients without requiring explicit characterization of post-selection distributions. Theoretically, we establish the asymptotic equivalence between post-selection inference targeting population-based regression coefficients and projection-based regression coefficients, the latter of which are the targets of existing post-selection inference methods, and prove the asymptotic validity of TwinPoSI in controlling the FCR. Extensive simulation studies demonstrate that TwinPoSI consistently outperforms existing approaches in FCR control, confidence interval length, and statistical power across diverse settings, including high-dimensional regimes, model misspecification, and weak-signal scenarios. I will further illustrate the practical utility of TwinPoSI through a real data application identifying significant biomarkers associated with labor timing and quantifying the strength of these associations.

DL20 — Mathematical Underpinnings of Distributed Inference: From Theory to Real-world Applications

June 15 (Monday) 11:00–12:40 **Room:** LT1B Program

Yong Chen, University of Pennsylvania, United States

Mathematical Underpinnings of Distributed Inference: From Theory to Real-world Applications

Multi-site studies are now central to biomedical research, but inference across sites remains challenging due to regulatory limits on sharing individual-level data, heterogeneous data distributions, sparse events in small centers, and the logistical burden of multi-round communication. In this talk, we introduce MOSAiC (Multi-site One-Shot Aggregation of Compressed Risk Functions), a unified framework for modern distributed research networks. Notably, MOSAiC reframes distributed learning as a mathematical problem of compressing and aggregating risk functions, leveraging recent advances in tensor networks—state-of-the-art tools for high-dimensional function approximation in scientific computing and computational physics. Our MOSAiC enjoys four desirable properties that have never been achieved by any of the existing federated algorithms (except linear regressions): one-shot communication, lossless recovery of pooled-data estimates, inclusiveness of all sites irrespective of size or event rarity, and analytic submodel exploitability without re-querying partners. We will illustrate MOSAiC's validity and efficiency through applications in drug relabeling, drug repurposing, and post-market safety surveillance.

Jingmei Qiu, University of Delaware, United States

TBD

TBD

Yudong Wang, National University of Singapore, Singapore

TT-MOSAiC: A Scalable Tensor-Train-Powered Framework for One-Shot and Lossless Federated Learning

This talk introduces TT-MOSAiC, a practical and scalable framework for one-shot and lossless federated learning. TT-MOSAiC leverages tensor train (TT) approximations to compress empirical loss functions at each data partner into low-rank tensor representations with manageable computational and storage costs. These compressed representations are then shared with a central server, which then recovers the local empirical losses of and performs downstream statistical inference. TT-MOSAiC requires only one round of communication of summary statistics, and can achieve arbitrarily close results to pooled analysis by adjusting tuning parameters. We will present the end-to-end workflow of TT-MOSAiC, including the construction of low-rank tensor representations via the TT-cross algorithm, the recovery of empirical loss functions using Chebyshev interpolation, and some key numerical and practical considerations for real-world deployment. We will also present its application to a decentralized study across four major U.S. healthcare systems involving approximately 120,000 patients, highlighting the effectiveness of TT-MOSAiC.

IP63 — Young Researchers' Session

June 15 (Monday) 11:00–12:40 **Room:** 209A Program

Minwoo Kim, Seoul National University, South Korea

Enhancing Differentially Private Mechanisms via Empirical Bayes

Differential privacy (DP) has emerged as the gold standard for privacy preservation in machine learning and statistical algorithms over the past few decades. While numerous methods have been developed to enhance the utility of DP algorithms under fixed privacy budgets, many of these approaches are computationally intensive or overly complex. In this presentation, we propose a novel approach that denoises the output of the standard additive Gaussian mechanism by leveraging empirical Bayes estimation. We demonstrate that the empirical Bayes approach significantly reduces mean-squared error by utilizing only the noisy output of the Gaussian mechanism. Our numerical studies illustrate that this simple yet powerful framework effectively improves performance across various statistical tasks—including histogram release, principal component analysis (PCA), and linear regression—often outperforming existing state-of-the-art private algorithms. Furthermore, we discuss extending this empirical Bayes approach to local DP settings.

Tetsuya Umino, University of Tsukuba, Japan

Automatic Sparse Estimation of High-Dimensional Mean Vectors in Multisample Settings

Scenarios involving high-dimensional, low-sample-size (HDLSS) data are often encountered in modern scientific fields, such as genomic analysis, where the number of variables greatly exceeds the number of observations. In multisample settings, reliable estimation of linear combinations of mean vectors is essential for characterizing the composite mean structure of multiple populations. However, classical estimators often exhibit severe noise accumulation. To address this issue, in this study, we propose a novel thresholding estimator of linear combinations of mean vectors for HDLSS settings. We consider the asymptotic properties of the conventional plug-in estimator and show that the estimator contains large amounts of noise in the high-dimensional setting, which renders it inconsistent. To solve this problem occurring in high-dimensional settings, we develop a new thresholding estimator based on the automatic sparse estimation methodology and show that the estimator is consistent under mild assumptions. We analyze and evaluate the performance of the proposed estimator based on numerical simulations and real data analysis. The simulations demonstrate that the method attains consistency without requiring the stringent high-dimensional conditions assumed by existing approaches. Furthermore, the real-data analysis illustrates its applicability to statistical problems involving mean structure inference, wherein improved estimation enhances the accuracy of the analysis.

Haruka Yoshida, Yokohama National University, Japan

Multiple Effect Restoration for Measurement and Confounding Bias in Causal Inference

In this talk, I consider the problem of identifying causal effects when the exposure, the outcome, or both are mismeasured in the presence of confounding bias that cannot be adequately controlled. Under the assumption that the causal relationships among variables can be represented by a directed acyclic graph and the corresponding recursive factorization of the joint distribution, I propose Multiple Effect Restoration (MER), which restores causal effects by leveraging multiple proxy variables that carry information about the mismeasured exposure, the mismeasured outcome, and/or unobserved confounders. MER extends existing proxy-based approaches, including Effect Restoration (Kuroki and Pearl, 2014) and Proximal Causal Learning (Miao et al., 2017), to settings with both measurement and confounding bias, and clarifies how the causal effects remain identifiable even in the presence of these biases.

Yuki Takazawa, The University of Tokyo, Japan

Robust Species Tree Inference Using Modes of Quartet Topologies in Tree Space

Quartet-based inference is a widely used approach to phylogenetic species tree estimation. It aggregates information from quartet trees, namely the unrooted trees obtained by restricting gene trees to sets of four taxa. Although such methods are often robust to incomplete lineage sorting, they can remain sensitive to irregular evolutionary processes such as horizontal gene transfer, gene flow, and hybridization.

In this work, we propose a mode-based quartet approach to species tree estimation. Each quartet tree observed across gene trees is treated as a random object in the one-dimensional Billera–Holmes–Vogtmann tree space, also known as the three-spider space, and is summarized by its mode rather than by empirical frequencies. This choice is motivated by the observation that frequency-based aggregation can be sensitive to contamination, whereas the mode provides a robust summary of the dominant phylogenetic signal in a contaminated distribution of quartet trees, downweighting atypical quartet trees arising from various sources of contamination. To make mode-based quartet aggregation computationally feasible for moderately large taxon sets, we also develop a heuristic search algorithm. Simulation studies indicate that the resulting estimator exhibits improved robustness compared to existing quartet-based methods.

IP67 — Advances in Flexible and Adaptive Statistical Inference

June 15 (Monday) 11:00–12:40 **Room:** 209B Program

Doudou Zhou, National University of Singapore, Singapore

MATES: Multi-view Aggregated Two-Sample Test

The two-sample test is a fundamental problem in statistics with a wide range of applications. In the realm of high-dimensional data, nonparametric methods have gained prominence due to their flexibility and minimal distributional assumptions. However, many existing methods tend to be more effective when the two distributions differ primarily in their first and/or second moments. In many real-world scenarios, distributional differences may arise in higher-order moments, rendering traditional methods less powerful. To address this limitation, we propose a novel framework to aggregate information from multiple moments to build a test statistic. Each moment is regarded as one view of the data and contributes to the detection of some specific type of discrepancy, thus allowing the test statistic to capture more complex distributional differences. The novel multi-view aggregated two-sample test (MATES) leverages a graph-based approach, where the test statistic is constructed from the weighted similarity graphs of the pooled sample. Under mild conditions on the multi-view weighted similarity graphs, we establish theoretical properties of MATES, including a distribution-free limiting distribution under the null hypothesis, which enables straightforward type-I error control. Extensive simulation studies demonstrate that MATES effectively distinguishes subtle differences between distributions. We further validate the method on the S&P100 data, showcasing its power in detecting complex distributional variations.

Kwangho Kim, Korea University, South Korea

Toward Flexible and Efficient Counterfactual Density Estimation

Comparing full counterfactual distributions provides richer measures of causal effects than conventional summaries such as means or quantiles. We study the problem of estimating the entire counterfactual density in a fully nonparametric setting and develop three complementary approaches. First, we introduce a doubly robust–style estimator that directly targets a kernel–smoothed counterfactual density. We establish its large-sample properties, derive finite-sample risk bounds, and construct uniform confidence bands via a bootstrap procedure. Second, we propose a diffusion- informed bump that adapts to the intrinsic geometry of the outcome manifold. By replacing the standard kernel with a diffusion-informed smoother, this estimator reduces bias near complex supports and attains faster convergence rates for high-dimensional outcomes when intrinsic dimension is low. Third, we develop a score-based method that targets the smoothed counterfactual score rather than the density itself. Focusing on the score enables even faster rates under common structural assumptions, with smaller constant factors, and can be paired with efficient density recovery when desired. We compare the three estimators and clarify when each is preferable, with particular emphasis on the advantages of diffusion-based smoothing for learning counterfactual distributions in high-dimensional settings. Together, these results provide a unified toolkit for flexible and statistically efficient counterfactual density estimation.

Jingru Zhang, Fudan University, China

Harmonizing Time-Varying Physical Activity Data Across Wearable Devices

Wearable devices and digital phenotyping have become central tools in observational and interventional studies for capturing real-time physical activity and other biosignals. Despite their growing availability, integrating and comparing wearable data across studies and cohorts remains challenging due to heterogeneity in device types, acquisition protocols, preprocessing pipelines, and measurement scales. These challenges are further compounded by the strong longitudinal and within-day correlations inherent in high-resolution time-varying sensor data.

In this talk, I will introduce INTACT (INtegration of Time-varying data from weArable sensors for physiCal acTivity), a novel statistical framework for harmonizing longitudinal physical activity intensity data collected from heterogeneous wearable devices. INTACT borrows strength across data sources by modeling shared eigenvalues and eigenfunctions, while allowing for source-specific scale and rotation adjustments, thereby effectively removing unwanted study- or device-specific effects without attenuating biologically meaningful variation. We illustrate the proposed method by integrating accelerometer data from two waves of the National Health and Nutrition Examination Survey (NHANES), which were collected using different devices and reported in different measurement units. INTACT effectively mitigates device- and study-specific effects while preserving biologically meaningful variation, allowing the integrated data to be analyzed jointly with increased effective sample size.

Jie Wang, The Chinese University of Hong Kong-Shenzhen, China

Variable Selection for Kernel Two-Sample Tests

We consider the variable selection problem for two-sample tests, aiming to select the most informative variables to determine whether two collections of samples follow the same distribution. To address this, we propose a novel framework based on the kernel maximum mean discrepancy (MMD). Our approach seeks a subset of variables with a pre-specified size that maximizes the variance-regularized kernel MMD statistic. We focus on three commonly used types of kernels: linear, quadratic, and Gaussian. From a computational perspective, we derive mixed-integer programming formulations and propose exact and approximation algorithms with performance guarantees to solve these formulations. From a statistical viewpoint, we derive the rate of testing power of our framework under appropriate conditions. These results show that the sample size requirements for the three kernels depend crucially on the number of selected variables, rather than the data dimension. Experimental results on synthetic and real datasets demonstrate the superior performance of our method, compared to other variable selection frameworks, particularly in high-dimensional settings.

IP72 — Bayesian Learning of Complex Structures

June 15 (Monday) 11:00–12:40 **Room:** 203 Program

Pierre Alquier, ESSEC Business School, Singapore

Rates of convergence in Bayesian meta-learning

The rate of convergence of Bayesian learning algorithms is determined by two conditions: the behavior of the loss function around the optimal parameter (Bernstein condition), the probability mass given by the prior to neighborhoods of the optimal parameter. In meta-learning, we face multiple learning tasks, that are independent but are still expected to be related in some way. For example, the optimal parameters of all the tasks can be close to each other. It is then tempting to use the past tasks to build a better prior, that we use to solve future tasks more efficiently. From a theoretical point of view, we hope to improve the prior mass condition in future tasks, and thus, the rate of convergence. In this paper, we prove that this is indeed the case. Interestingly, we also prove that we can learn the optimal prior at a fast rate of convergence, regardless of the rate of convergence within the tasks (in other words, Bernstein condition is always satisfied for learning the prior, even when it is not satisfied within tasks).

This is a joint work with Badr-Eddine Chérief-Abdellatif (CNRS) and Charles Riou (RIKEN AIP and The University of Tokyo).

Jaeyong Lee, Seoul National University, Korea

Eigenstructure inference for the high-dimensional covariance matrices with generalized shrinkage inverse- Wishart prior

In high-dimensional settings, where the number of covariates increases with the sample size, it is well known that the eigenstructure of the sample covariance matrix is inconsistent. The inverse-Wishart prior, a standard choice for covariance estimation in Bayesian inference, also suffers from posterior inconsistency. To address the issue of eigenvalue dispersion in high-dimensional settings, the shrinkage inverse-Wishart (SIW) prior has recently been proposed. Despite its conceptual appeal and empirical success, the asymptotic justification for the SIW prior has remained limited. In this paper, we propose a generalized shrinkage inverse-Wishart (gSIW) prior for high-dimensional covariance modeling. By extending the SIW framework, the gSIW prior accommodates a broader class of prior distributions and facilitates the derivation of theoretical properties under specific parameter choices. In particular, under the spiked covariance assumption, we establish the asymptotic behavior of the posterior distribution for both eigenvalues and eigenvectors by directly evaluating the posterior expectations for two sets of parameter choices. This direct evaluation provides insights into the large-sample behavior of the posterior that cannot be obtained through general posterior asymptotic theorems. Finally, simulation studies illustrate that the proposed prior provides accurate estimation of the eigenstructure, particularly for spiked eigenvalues, achieving narrower credible intervals and higher coverage probabilities compared to existing methods.

William Weimin Yoo, Heriot-Watt University Malaysia, Malaysia

Learning Weights and Depth in Bayesian Neural Networks via Markov Chain Approximations

In this talk, we will look at the problem of efficiently learning the network weights and biases of Bayesian neural networks (BNNs). By endowing Gaussian priors on the network parameters, we will obtain posterior distribution of the BNN.

However, simulating from this posterior can be prohibitively expensive due to the complex compositional and nonlinear structure of BNNs. We propose to approximate this posterior by constructing a Markov chain across the layers and the parameters are learned via variational inference. By sampling outputs from the BNN Markov sampler, we study the related problems of learning the network depth and credible set construction for uncertainty quantification. In addition, we will discuss differences between the proposed method with other Bayesian approaches in the literature such as approaches based on sparsity inducing and heavy-tailed priors.

Subhashis Ghoshal, North Carolina State University, United States

Bayesian learning of relational graph in semiparametric high-dimensional time series

Time series data arising in many applications nowadays are high-dimensional. A large number of parameters describe features of these time series. Sensible inferences on these parameters with limited data are possible if some underlying lower-dimensional structure is present. We propose a novel approach to modeling a high-dimensional time series through several independent univariate time series, which are then orthogonally rotated and sparsely linearly transformed. With this approach, any specified intrinsic relations among component time series, as given by a graphical structure, can be maintained at all time snapshots. We call the resulting process an Orthogonally Rotated Univariate Time Series (OUT). Key structural properties of time series, such as stationarity and causality, can be easily accommodated in the OUT model. For Bayesian inference, we put suitable prior distributions on the spectral densities of the independent latent time series, the orthogonal rotation matrix, and the common precision matrix of the component time series at every time point. A likelihood is constructed using the Whittle approximation for univariate latent time series. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. We study the convergence of the pseudo-posterior distribution based on the Whittle likelihood for the model's parameters upon developing a new general posterior convergence theorem for pseudo-posteriors. We find that the posterior contraction rate for independent observations essentially prevails in the OUT model under very mild conditions on the temporal dependence described in terms of the smoothness of the corresponding spectral densities. In establishing the result, we develop a new general theorem on the contraction rate of a pseudo-posterior distribution, which is potentially applicable in other settings. Through a simulation study, we compare the accuracy of parameter estimation and graphical structure identification with other approaches. We apply the proposed methodology to analyze a dataset on various industrial components of the US gross domestic product from 2010 to 2019 and to predict future observations.

Based on a collaboration with Arkaprava Roy, University of Florida, and Anindya Roy, University of Maryland-Baltimore County.

IP39 — Innovative Methods in Survival Analysis and Survey Data Integration

June 15 (Monday) 11:00–12:40 **Room:** 201 Program

Jianguo Sun, Southern University of Science and Technology, China

A new transfer learning estimation approach for failure time data

In this talk, we discuss transfer learning for regression analysis of interval-censored failure time data. Some new estimation methods will be presented and shown to be effective. In addition, they will be applied to a motivating example.

Dipankar Bandyopadhyay, Viginia Commonwealth University, United States

Rank estimation for the accelerated failure time model under partially interval-censored data

In this talk, we present a unified rank-based inferential procedure for fitting the accelerated failure time model to partially interval-censored (PIC) data. A Gehan-type monotone estimating function is constructed based on the idea of the familiar weighted log-rank test, and an extension to a general class of rank-based estimating functions are suggested. The proposed estimators can be obtained via linear programming, and are shown to be consistent and asymptotically normal via standard empirical process theory. Unlike common maximum likelihood-based estimators for PIC regression models, our approach can directly provide a regression coefficient estimator without involving a complex nonparametric estimation of the underlying residual distribution function. An efficient variance estimation procedure for the regression coefficient estimator is considered. Moreover, we extend the proposed rank-based procedure to the linear regression analysis of multivariate cluster-correlated PIC data. The finite-sample operating characteristics of our approach are examined via simulation studies. Data analysis from a colorectal cancer study illustrates the practical utility of the method.

This is joint work with Drs. Sangbum Choi, and Taehwa Choi.

Changbao Wu, University of Waterloo, Canada

Data Integration with Non-Probability Survey Samples

Non-probability survey samples have become a rich source of information in the big data era for scientific investigations in many fields, including medical studies and public health research. In this presentation, we first provide a brief review of recent methodological developments on data integration with non-probability survey samples for valid and efficient statistical inference. We then describe an application of the inverse probability weighting (IPW) method and the integrated poststratification strategy to the non-probability survey conducted by the Canadian Institute for Health Information (CIHI) on measuring access to home and community care and to mental health and substance use services in Canada.

Yi Li, University of Michigan, United States

Inference for the Relative Risk Functional in Deep Nonparametric Cox Models

There remain theoretical gaps in deep neural network estimators for the nonparametric Cox proportional hazards model. In particular, it is unclear how gradient-based optimization error propagates to population risk under partial likelihood, how pointwise bias can be controlled to permit valid inference, and how ensemble-based uncertainty quantification behaves under realistic variance decay regimes.

We develop an asymptotic distribution theory for deep Cox estimators that addresses these issues. First, we establish nonasymptotic oracle inequalities for general trained networks that link in-sample optimization error to population risk without requiring the {exact} empirical risk {optimizer}. We then construct a structured neural parameterization that achieves

infinity-norm approximation rates compatible with the oracle bound,

yielding control of the pointwise bias.

Under these conditions and using the H{a}jek–Hoeffding projection, we prove pointwise and multivariate asymptotic normality for subsampled ensemble estimators. We derive a range of subsample sizes that balances bias correction with the requirement that the H{a}jek–Hoeffding projection remain dominant. This range accommodates decay conditions on the single-overlap covariance, which measures how strongly a single shared observation influences the estimator, and is weaker than those imposed in the subsampling literature.

An infinitesimal jackknife representation provides analytic covariance estimation and valid Wald-type inference for relative risk contrasts such as log-hazard ratios. Finally, we illustrate the finite-sample implications of the theory through simulations and a real data application.

IP66 — Statistical Methods for Complex and Non-Euclidean Data

June 15 (Monday) 11:00–12:40 **Room:** 202 Program

Ho Yun, École Polytechnique Fédérale de Lausanne, Switzerland

Spectral Gaps and Spherical Harmonics: A Directional Statistics Approach to DNA Flexibility

Understanding the mechanical properties of biopolymers, such as DNA and protein filaments, requires robust mathematical frameworks. The wormlike chain model has long served as a cornerstone in structural biology for this purpose, relying on "persistence length" to quantify bending stiffness. In this session, we re-examine this fundamental parameter through the lens of spherical harmonics, framing persistence length as the spectral gap of an underlying diffusion operator. This perspective bridges structural mechanics with directional statistics, naturally giving rise to classic conformational models like the von Mises-Fisher distribution. Finally, we will explore how these insights inspire generalized statistical models for analyzing DNA sequential data, highlighting the rich, interdisciplinary intersection of molecular biology and modern statistics.

Almond Stöcker, École Polytechnique Fédérale de Lausanne, Switzerland

Kernel ridge regression for spherical responses

The aim is to propose a novel nonlinear regression framework for responses taking values on a hypersphere. Rather than performing tangent space regression, where all the sphere responses are lifted to a single tangent space on which the regression is performed, we estimate conditional Frechet means by minimizing squared distances on the nonlinear manifold. Yet, the tangent space serves as a linear predictor space where the regression function takes values. The framework integrates Riemannian geometry techniques with functional data analysis by modelling the regression function using methods from vector-valued reproducing kernel Hilbert space theory. This formulation enables the reduction of the infinite-dimensional estimation problem to a finite-dimensional one via a representer theorem and leads to an estimation algorithm by means of Riemannian gradient descent. Explicit checkable conditions on the data that ensure the existence and uniqueness of the minimizing estimator are given.

Seungwoo Kang, Sungkyunkwan University, South Korea

L_1 Prominence Measures for Directed Graphs

We introduce novel measures, L_1 prestige and L_1 centrality, for quantifying the prominence of each vertex in a strongly connected and directed graph by utilizing the concept of L_1 data depth (Vardi and Zhang, Proc. Natl. Acad. Sci. U.S.A. 97(4):1423–1426, 2000). The former measure quantifies the degree of prominence of each vertex in receiving choices, whereas the latter measure evaluates the degree of importance in giving choices. The proposed measures can handle graphs with both edge and vertex weights, as well as undirected graphs. However, examining a graph using a measure defined over a single 'scale' inevitably leads to a loss of information, as each vertex may exhibit distinct structural characteristics at different levels of locality. To this end, we further develop local versions of the proposed measures with a tunable locality parameter. Using these tools, we present a multiscale network analysis framework that provides much richer structural information about each vertex than a single-scale inspection. By applying the proposed measures to the networks constructed from the Seoul Mobility Flow Data, it is demonstrated that these measures accurately depict and uncover the inherent characteristics of individual city regions.

Seoncheol Park, Hanyang University, South Korea

Adaptive Boosting on Linear Networks

Classification is a supervised machine learning method that predicts a categorical response variable using several explanatory variables. If observations are sampled from a spatial point process, then we can also use x- and y-coordinates as explanatory variables. If the observations are sampled from a known linear network instead of whole space, then the distance between two points is defined differently, and we require a classifier for the linearly clustered data. In this study, we address the classification problem on a tree-shaped linear network. We select a point on the edges in the given linear network to split the space, and then construct a decision tree through recursive splits. We propose an adaptive boosting algorithm using this decision tree as a weak classifier. Finally, we provide some simulated examples and real data analysis, comparing with adaptive boosting based on decision trees constructed using Cartesian coordinates. The proposed method has better accuracy than the comparison method, when the observations are clustered on linear network. This work was supported by the research fund of Hanyang University (HY-202300000001150) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00339064).

CS01 — Bayesian and Likelihood-Based Inference

June 15 (Monday) 11:00–12:40 **Room:** 214 Program

Weichang Yu, University of Melbourne

Cutting Feedback in Misspecified Copula Models

In copula models the marginal distributions and copula function are specified separately. We treat these as two modules in a modular Bayesian inference framework, and propose conducting modified Bayesian inference by "cutting feedback". Cutting feedback limits the influence of potentially misspecified modules in posterior inference. We consider two types of cuts. The first limits the influence of a misspecified copula on inference for the marginals, which is a Bayesian analogue of the popular Inference for Margins (IFM) estimator. The second limits the influence of misspecified marginals on inference for the copula parameters by using a pseudo likelihood of the ranks to define the cut model. We establish that if only one of the modules is misspecified, then the appropriate cut posterior gives accurate uncertainty quantification asymptotically for the parameters in the other module. Computation of the cut posteriors is difficult, and new variational inference methods to do so are proposed. The efficacy of the new methodology is demonstrated using both simulated data and a substantive multivariate time series copula application from macroeconomic forecasting. In the latter, cutting feedback from misspecified marginals to a 1096 dimension copula improves posterior inference and predictive accuracy greatly, compared to conventional Bayesian inference. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Yichen Zhu, The University of Hong Kong

Vecchia Gaussian Processes: On Probabilistic and Statistical Properties

Gaussian Processes (GPs) are widely used to model dependencies in spatial statistics and machine learning. However, exact inference is computationally intractable for GP regression, with a time complexity of O(n^3). The Vecchia approximation scales up computation by introducing sparsity into the spatial dependency structure, represented by a directed acyclic graph (DAG). Despite its practical popularity, this approach lacks rigorous theoretical foundations, and the choice of DAG structure remains an open problem.

In this paper, we systematically study the Vecchia approximation of the popular, isotropic Matérn GP as standalone stochastic process and uncover key probabilistic and statistical properties. We propose selecting parent sets as norming sets with fixed cardinality in the Vecchia approximation. On the probabilistic side, we show that the conditional distributions of Matérn GPs, as well as their Vecchia approximations, can be characterized by polynomial interpolations. This enables us to establish several results on small ball probabilities and the Reproducing Kernel Hilbert Spaces (RKHSs) of Vecchia GPs. Building on these probabilistic results, we prove that in the nonparametric regression model, the corresponding posterior contracts around the truth at the optimal minimax rate, both under oracle rescaling and hierarchical tuning of the prior.

We illustrate the theoretical findings through numerical experiments on synthetic datasets. Our core algorithms are implemented in C++ with an R interface.

Khue-Dung Dang, University of Western Australia

Variational Approximate Penalized Credible Regions for Bayesian Grouped Regression

We develop a fast and accurate grouped penalized credible region approach for variable selection and prediction in Bayesian high-dimensional linear regression. Most existing Bayesian methods either are subject to high computational costs due to long Markov Chain Monte Carlo runs or yield ambiguous variable selection results due to non-sparse solution output. On the other hand, the penalized credible region framework yields sparse post-processed estimates that facilitates unambiguous grouped variable selection. We propose using a grouped global-local shrinkage prior to achieve high estimation accuracy and approximate posterior summaries using coordinate ascent variational inference. The penalized credible region framework is then recast as a convex optimization problem that admits efficient computations. We prove that the resultant post-processed estimators are both parameter-consistent and variable selection consistent in high-dimensional settings. Theory is developed to justify running the coordinate ascent algorithm for at least two cycles. Through extensive simulations, we demonstrate that our proposed method outperforms state-of-the-art methods in grouped variable selection, prediction, and computation time for several common models including ANOVA and nonparametric varying coefficient models.

Nelson Jinn-Yih Chua, Australian National University

Asymptotic Results for Model Parameters and Random Effects Inference Using Gaussian Variational Approximations in Generalized Linear Mixed Models

In recent years, Gaussian variational approximations (GVA) have become a popular approach for frequentist inference in models involving unobserved or latent random variables, providing an alternative objective function to the marginal log-

likelihood that is computationally more efficient to optimize. GVA also produces predictions and associated uncertainty measures for the unobserved random effects, based directly on the mean and (co)variance parameters of Gaussian variational distribution, respectively. Despite these computational benefits, the statistical properties of GVA remain relatively underexplored in the literature, owing to the lack of large-sample results that are applicable to, say, a range of response types.

In this talk, we present a set of asymptotic properties for the parameter estimates and random effects predictions obtained from GVA in the setting of independent-cluster generalized linear mixed models. We show that for model parameter estimation such as fixed effect coefficients and variance components, GVA is asymptotically fully efficient relative to maximum likelihood estimation. For random effects inference, we demonstrate that the uncertainty measure produced by GVA is inadequate in certain sample size settings, and we propose a simple mathematical adjustment to rectify this. Properties such as convergence rates are also shown to depend substantially on whether or not the true random effects are conditioned on, and we discuss the practical implications of this.

Shogo Kusano, Kumamoto University

Quasi-Bayesian Information Criterion of SEM for Diffusion Processes

Structural equation modeling (SEM) is a statistical method for examining relationships among unobservable variables. As a confirmatory analysis method, SEM requires the model to be specified in advance based on the theoretical framework of the respective research field. However, in practice, statisticians often have several candidate SEMs and need to select the most appropriate one among them. To address this issue, a variety of information criteria for SEM have been developed. In recent years, SEM for diffusion processes based on high-frequency data has been proposed, and a quasi-Akaike information criterion for the SEM has also been studied. However, this criterion does not guarantee model selection consistency. In this study, based on an asymptotic expansion of the marginal quasi-log-likelihood, we propose two types of quasi-Bayesian information criteria for the SEM, and show that the proposed criteria have model selection consistency. Several examples and simulation results are also presented.

CS13 — Applied Probability, Financial Statistics, and Cure Models

June 15 (Monday) 11:00–12:40 **Room:** 215 Program

Hugh Entwistle, Macquarie University

On Optimal Stopping Problems with Random Supply

The classic 'best choice problem' where one wishes to maximise their probability of selecting the best, with only one choice, when observing a random permutation of objects sequentially is a well-studied problem. Two additional variants of this problem are when the number of objects is random and; when the objective is to instead maximise the probability of selecting the best and second best with two choices. With a short journey and history of these problems, we turn to investigate the variant where both of these assumptions hold - we are faced with a random number of observations and wish to stop on the best and second best observation. We specifically explore the structure of the optimal rule and show-case some interesting results based on the closure of the stopping sets. Time permitting, we may discuss the role of random supply in the full information variant.

Yuanhang Luo, The Hong Kong Polytechnic University

Adaptive Debiased Lasso in High-dimensional Generalized Linear Models with Streaming Data

Online statistical inference facilitates real-time analysis of sequentially collected data, making it different from traditional methods that rely on static datasets. This paper introduces a novel approach to online inference in high-dimensional generalized linear models, where we update regression coefficient estimates and their standard errors upon each new data arrival. In contrast to existing methods that either require full dataset access or large-dimensional summary statistics storage, our method operates in a single-pass mode, significantly reducing both time and space complexity. The core of our

methodological innovation lies in an adaptive stochastic gradient descent algorithm tailored for dynamic objective functions, coupled with a novel online debiasing procedure. This allows us to maintain low-dimensional summary statistics while effectively controlling the optimization error introduced by the dynamically changing loss functions. We establish the asymptotic normality of our proposed Adaptive Debiased Lasso (ADL) estimator. We conduct extensive simulation experiments to show the statistical validity and computational efficiency of our ADL estimator across various settings. Its computational efficiency is further demonstrated via a real data application to the spam email classification.

Lovely Aisha Jamil, American University of Sharjah

Understanding the Evolution of Kyle's Lambda on Digital Blockchain Assets

Kyle's lambda measures the price impact of trading activity on an asset and serves as a proxy for market liquidity. While extensively studied in traditional financial markets, its dynamism in digital asset markets remains underexplored. This paper examines the evolution of Kyle's lambda across major blockchain-based assets, including cryptocurrencies, decentralized finance (DeFi) tokens, non-fungible tokens, meme coins, and fan tokens, to analyze how liquidity sensitivity varies across different assets, currencies, time frames, and market conditions. The results show heterogeneity in Kyle's lambda across digital asset classes, with cryptocurrencies and large-cap DeFi tokens exhibiting lower liquidity sensitivity than NFTs and meme coins. Kyle's lambda is negatively related to dollar trading volume ($p < 0.01$), positively associated with return volatility, spikes during market stress, and dynamically precedes increases in volatility and declines in trading activity, indicating that it serves as an early signal of liquidity risk and varying market efficiency across blockchains. These findings contribute to the literature on digital market microstructure by evaluating cross-sectional differences in price impact and demonstrating that liquidity measures contain forward-looking information about market fragility. These results provide relevant implications for traders, researchers, and regulators concerned with liquidity risk and the stability of decentralised financial systems.

Junyan Ye, The Chinese University of Hong Kong

Martingale Duality Meets Statistical Learning: Deep Primal-Dual Bounds and Policy Learning for High-Dimensional Optimal Switching Problem

We connect martingale duality and statistical learning for finite-horizon optimal switching with discrete intervention dates on a general filtration, allowing continuous-time observations between decision times. We derive a martingale dual representation for multiple switching, where the minimal penalty is characterized by the Doob martingales of continuation values, yielding fully computable upper bounds. Extending DeepMartingale from optimal stopping to optimal switching, we prove convergence under both the upper-bound loss and an L^2-surrogate loss. We also establish a dimension-robust expressivity result: for any $\varepsilon > 0$, neural networks of size $cd^q\varepsilon^{-r}$, with constants independent of d and ε, achieve ε-accurate dual upper bounds, thereby avoiding the curse of dimensionality. Our theory further suggests dimension-scaling laws for architecture design, training, and hedging rebalancing. Beyond dual upper bounds, we develop a dual-guided statistical learning approach for recovering an adapted and interpretable primal switching policy, together with a genuine lower bound. We further provide theoretical guarantees for the resulting primal learning. Numerical experiments on Brownian and Brownian–Poisson models show small primal–dual gaps and strong high-dimensional performance.

DL13 — Discrete Random Models and Learning

June 16 (Tuesday) 09:00–10:40 **Room:** LT1A Program

Antar Bandyopadhyay, Indian Statistical Institute, Delhi, India

Interacting Urn Schemes

In this talk, we will introduce a novel class of positive reinforcement models which we will refer to as "interactive urn schemes" with the goal of obtaining a limiting distribution. As we will see these set of models have resemblance with examples of "Self-Organized Criticality (SOC)". The interactions will be defined via a network (possibly infinite). We will

show that limit exists under fairly general condition and for all type of networks as long as it is "locally finite". We will further indicate a new technique which is used to prove for more general structure including undirected graphs.

[This is a joint work with Deborshi Das]

Subhro Ghosh, National University of Singapore, Singapore

Strongly correlated particle systems: a toolbox for machine intelligence

The classical paradigm of randomness in the sciences is that of i.i.d. random variables, and going beyond i.i.d. is often considered a difficulty and a challenge to be overcome. In this talk, we will explore a new perspective, wherein strongly constrained random systems in fact help to understand fundamental problems in machine learning. In particular, we will discuss strongly correlated particle systems that are well-motivated from statistical and quantum physics, including in particular determinantal probability measures. These will be used to shed important light on questions of fundamental interest in learning theory, focussing on applications to novel sampling techniques and advances in stochastic gradient descent.

Based in part on the following works:

[1] Gaussian determinantal processes: A new model for directionality in data, (with P. Rigollet) [Proceedings of the National Academy of Sciences (PNAS), vol. 117, no. 24, pp. 13207–13213]

[2] Small coresets via negative dependence: DPPs, linear statistics, and concentration, (with R. Bardenet, H. Simon-Onfroy, H.S. Tran) [Spotlight at NeurIPS 2024]

[3] Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD, (with R. Bardenet, M. Lin) [Spotlight at NeurIPS 2021]

[4] Negative Dependence as a toolbox for machine learning: review and new developments, (with H.S. Tran, V. Petrovic, R. Bardenet) [arxiv.org/abs/2502.07285]

Nathan Ross, University of Melbourne, Australia

Detecting correlation in uniform attachment trees

We introduce and study a new model of correlated uniform attachment (UA) trees, where correlation is sprinkled throughout the time evolution of the process. In this model, two UA trees are grown in parallel from a single node labeled 0, and at the nth time step, a new node labeled n is added to each tree, with an edge between it and a uniformly chosen existing vertex in the respective tree. The two choices of attachment are correlated: With probability α, the edges attach to nodes with the same label in both trees, and with probability 1-α, the choices are made independently.

A fundamental question is how well the correlation can be detected from a single pair of unlabeled correlated UA trees. We show that this can be done with probability tending to one as the size of the trees goes to infinity, by constructing a statistic of the unlabeled trees that converges to the correlation parameter α.

The construction of our statistic relies on two key ideas. The first is that we can use a notion of centrality to identify subsets of vertices of each tree whose intersection has a sufficient number of common early vertices. The second idea is that across different scales, it is possible to approximately determine the labels of vertices that have attached to these early vertices, using the sizes of fringe subtrees. Our analysis includes quantitative bounds on the fraction of early vertices that remain most central, which may be of independent interest. Joint work with Johannes Bäumler, Miklós Rácz, and Anirudh Sridhar.

DL14 — Development of Causal Inference

June 16 (Tuesday) 09:00–10:40 **Room:** LT1B Program

Manabu Kuroki, Yokohama National University, Japan

The Evaluation of the Probablities of Potential Outcome Types from Statistical Data

The concept of potential outcome types is one of the fundamental components of causal inference. Even in randomized experiments, existing researches have shown that the probabilities of potential outcome types are generally not identifiable,

leading to bounds such as the Tian−Pearl bounds (Tian and Pearl, 2000), unless additional assumptions such as monotonicity are imposed. In this talk, based on a series of joint studies with my colleagues (Kawakami, Shingaki, and Yoshida, 2021−2025), I introduce novel identification conditions for the probabilities of potential outcome types that do not rely on monotonicity or other strong structural assumptions, by exploiting information from proxy covariates.

Yuta Kawakami, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

Moments of Causal Effects

The moments of random variables are fundamental statistical measures for characterizing the shape of a probability distribution, encompassing metrics such as mean, variance, skewness, and kurtosis. On the other hand, the primary focus of causal inference is the evaluation of causal effects, which are defined as the difference between two potential outcomes. While traditional causal effect assessment focuses on the average causal effect, I explain moments of causal effects in this talk to analyze the distribution of causal effects.

Shunichiro Orihara, Tokyo Medical University, Japan

Average Treatment Effect Estimation under Poor Overlap via Weighted Estimands

We commonly consider the average treatment effect (ATE) as the causal effect of interest. In situations with poor overlap, where the covariate support between the treatment and control groups is limited, the inverse probability weighting (IPW) estimator may have large variance. To address this issue, weighted ATE (WATE) estimands, such as the ATE for the overlap population (ATO), are sometimes considered. However, these estimands may differ from the original target of interest such as the ATE. In this presentation, we propose a novel estimation procedure that uses a class of WATEs based on the beta weight family, including ATO, and estimates the ATE via extrapolation. Specifically, we consider the following steps: 1) estimating IPW estimators for several WATE estimands; 2) regressing the IPW estimates on the hyperparameter of the WATEs; and 3) setting the hyperparameter to 0 (extrapolation). We discuss theoretical justifications for the procedure and its properties, such as asymptotic normality. Additionally, since the proposed procedure is easy to implement, we will explain how to conduct the analysis in R with a simple programming.

IP29 — Recent Advances in Reinforcement Learning

June 16 (Tuesday) 09:00−10:40 **Room:** 209A Program

Daniele Bracale, University of Michigan, United States

Online Price Competition under Generalized Linear Demands

We study sequential price competition among N sellers, each influenced by the pricing decisions of their rivals. Specifically, the demand function for each seller i follows the single index model $\lambda_i(\mathbf{p}) = \mu_i(\langle \boldsymbol{\theta}_{i,0}, \mathbf{p} \rangle)$, with known increasing link μ_i and unknown parameter $\boldsymbol{\theta}_{i,0}$, where the vector $\mathbf{p}$ denotes the vector of prices offered by all the sellers simultaneously at a given instant. Each seller observes only their own realized demand – unobservable to competitors – and the prices set by rivals. Our framework generalizes existing approaches that focus solely on linear demand models. We propose a novel decentralized policy, PML-GLUCB, that combines penalized MLE with an upper-confidence pricing rule, removing the need for coordinated exploration phases across sellers – which is integral to previous linear models – and accommodating both binary and real-valued demand observations. Relative to a dynamic benchmark policy, each seller achieves $O(N^2\sqrt{T}\log(T))$ regret, which essentially matches the optimal rate known in the linear setting. A significant technical contribution of our work is the development of a variant of the elliptical potential lemma – typically applied in single-agent systems – adapted to our competitive multi-agent environment.

Kihyuk Hong, Korea Advanced Institute of Science and Technology, South Korea

Recent Advances in Offline Constrained Reinforcement Learning

Offline reinforcement learning is a framework for learning decision-making policies entirely from previously collected data, without further interaction with the environment. This is particularly important in applications such as robotics, healthcare, and recommendation systems, where trial-and-error interaction can be costly, slow, or unsafe. In many such applications, however, performance alone is not enough: learned policies must also satisfy safety, resource, fairness, or operational constraints. This motivates the study of offline constrained reinforcement learning, where the goal is to learn a high-performing policy from logged data while respecting prescribed constraints.

In recent years, offline reinforcement learning has achieved significant practical success across a wide range of applications. Despite this progress, the theoretical foundations needed to understand when and why offline reinforcement learning works remain more limited, especially in the constrained setting. Since 2020, a growing body of research has made substantial progress in addressing these questions, clarifying the roles of data coverage, function approximation, pessimism, and uncertainty quantification. In this talk, I will provide an overview of these theoretical developments, discuss the central challenges of offline reinforcement learning and offline constrained reinforcement learning, and highlight several recent advances that deepen our understanding of when reliable policy learning from static data is possible.

Shivaram Kalyanakrishnan, Indian Institute of Technology Bombay, India

On-line Learning in Tree MDPs by Treating Policies as Bandit Arms

A Tree Markov Decision Problem (T-MDP) is a finite-horizon MDP with a starting state s1, in which every state is reachable from s1 through exactly one state-action trajectory. T-MDPs arise naturally as abstractions of decision making in sequential games with perfect recall, against stationary opponents. We consider the problem of on-line learning in T-MDPs, both in the PAC and the regret-minimisation regimes. We show that well-known bandit algorithms—Lucb and Ucb—can be applied on T-MDPs by treating each policy as an arm. The apparent technical challenge in this approach is that the number of policies is exponential in the number of states. Our main innovation is in the design of confidence bounds based on data shared by the policies, so that the bandit algorithms can yet be implemented with polynomial memory and per-step computation. We obtain instance-dependent upper bounds on sample complexity and regret that sum a "gap term" from every terminal state, rather than every policy. Empirically, our algorithms consistently outperform available alternatives on a suite of hidden-information games.

Wen Sun, Cornell University, United States

TBD

TBD

IP12 — Experiment Design in the Modern Era

June 16 (Tuesday) 09:00–10:40 **Room:** 209B Program

Lixin Zhang, Zhejiang GongShang University, China

Asymptotic Properties of Covariate Adaptive Randomization Procedures for Balancing Observed and Unobserved Covariates

Balancing treatment allocation for influential covariates has become increasingly important in today's clinical trials. Covariate adaptive randomization (CAR) procedures are extensively used to reduce the likelihood of covariate imbalances occurring. In this talk, we consider a general framework of CAR procedures for clinal trials which can balance general covariate features, such as quadratic and interaction terms which can be discrete, continuous, and mixing. We show that under widely satisfied conditions the proposed procedures have superior balancing properties; in particular, the convergence rate of imbalance vectors can attain the best rate $O_P(1)$, and at the same time, the convergence rate of the imbalance of unobserved covariates is $O_P(\sqrt{n})$, where n is the sample size. The variance inflation problem and shift problem of the covariate adaptive randomization procedures are considered and new covariate adaptive randomization procedures without the variance inflation problem and shift problem are proposed.

Wei Ma, Renmin University of China, China

Covariate-adaptive design: An overview and recent advances

Covariate-adaptive designs are a class of experimental design methods that dynamically adjust treatment allocation probabilities to achieve balanced covariates across treatment groups. Because of their strengths in enhancing treatment group comparability, increasing the precision of treatment effect estimation, and producing more convincing experimental results, these designs are extensively employed in randomized controlled settings, including clinical trials, economic field experiments, and online A/B testing. This talk first provides a methodological review of various covariate-adaptive design approaches and then discusses a recent advancement in the field, which proposes a novel and unified framework for covariate-adaptive designs. The challenges and solutions in analyzing data collected from covariate-adaptive designs will also be addressed.

Waverly Wei, University of Southern California, United States

Trace-Aware Routing for Cost-Effective Human–AI Collaborative Labeling

Large-scale generative AI systems, such as text-to-image models, require reliable labels to evaluate whether outputs satisfy intended specifications (e.g., prompt fidelity or rubric-based quality). In practice, AI labelers (e.g., large language model (LLM)/vision-language model (VLM) judges) are efficient but may exhibit systematic errors on subtle or fine-grained aspects of text–image alignment, whereas human labels can be more reliable but costly. This raises a natural question: How can we coordinate AI and human labelers so that only instances likely to be mislabeled are escalated to humans, under a fixed human-labeling budget? Furthermore, in many modern AI labeling workflows that rely on LLM/VLM judges, labels are produced together with an explicit reasoning trace prior to finalization, allowing early human intervention and potential savings in AI labeling computational cost. Yet, a key challenge is *when to escalate*, as intervening too late wastes AI computation and may fail to prevent incorrect labels, while intervening too early incurs unnecessary human effort. Existing deferral and routing methods typically make one-shot decisions and do not exploit trace information. Here, we construct a trace-aware AI router for cost-effective human–AI collaboration with three key features: (i) it conditions routing decisions on the evolving reasoning trace of the AI labeler; (ii) it performs stepwise monitoring to determine the earliest point at which human review is needed; and (iii) it incorporates human budget control through a feature-based disagreement scoring model, prioritizing hard instances where AI and human judgments are more likely to differ. Empirical results across multiple benchmarks and baselines show that our method consistently improves labeling accuracy under fixed human budgets, demonstrating the value of reasoning traces for sequential, budget-aware routing.

Jingshen Wang, University of California, Berkeley, United States

TBD

TBD

IP21 — Innovations in Changepoint Detection

June 16 (Tuesday) 09:00–10:40 **Room:** 203 Program

Ying Yang, Fudan University, China

Spatially Randomized Designs Can Enhance Policy Evaluation

This work 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.

Wei Zhang, Fudan University, China

Structural Change Detection in Dynamic Systems

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 (ODEs). Unlike existing methods that assume simple 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.

Chengde Qian, Shanghai Jiao Tong University, China

Changepoint Detection in Complex Models: Cross-Fitting Is Needed

Changepoint detection is commonly formulated by minimizing the sum of in-sample losses to quantify the model's overall fit. However, for flexible modeling procedures—especially those involving high-dimensional parameter spaces or hyperparameter tuning—this strategy can lead to inaccurate changepoint estimation due to over-adaptivity biases. To mitigate this issue, we propose a novel cross-fitting methodology based on out-of-sample loss evaluations, which decouples model fitting from changepoint search. We establish a general theoretical framework for consistent changepoint estimation under mild conditions, and further extend it to temporally dependent data. A key implication of the theory is that consistency depends primarily on the models' predictive accuracy over nearly homogeneous segments. Numerical experiments show that the proposed method substantially improves the reliability and adaptability of changepoint detection in complex scenarios.

Guanghui Wang, Nankai University, China

ART: Distribution-Free and Model-Agnostic Changepoint Detection with Finite-Sample Guarantees

We introduce ART, a distribution-free and model-agnostic framework for changepoint analysis with finite-sample guarantees. ART transforms independent observations into real-valued scores via a symmetric function; under the null hypothesis of no changepoint these scores are exchangeable. Ranking and aggregating the scores yields test statistics whose null distribution is known exactly from the permutation law of ranks, enabling exact finite-sample Type I error control without repeated refitting under permutations. ART extends naturally to a multi-scale setting: by locally ranking scores over a family of intervals and aggregating them, it supports multiple changepoint testing, localization with inference, and post-detection inference, while retaining distribution-free calibration. The approach is model-agnostic: it imposes minimal structural or distributional assumptions and accommodates diverse score constructions, including features learned by statistical or machine-learning models. Across simulations and real-data applications, ART delivers valid error control and competitive power across a range of models and distributions. These properties make ART a reliable and versatile tool for modern changepoint analysis.

IP52 — Recent Advancements in Semiparametric Methods for Complex Censored Outcomes

June 16 (Tuesday) 09:00–10:40 **Room:** 201 Program

Yu Gu, The University of Hong Kong, Hong Kong

Semiparametric Functional Multi-State Models with Application to Alzheimer's Disease

Recent advances in brain imaging have greatly enhanced early and accurate prediction of Alzheimer's disease (AD) risk. However, existing imaging-based methods focus solely on the progression from mild cognitive impairment (MCI) to AD, and improperly assume that the exact time of AD conversion is known or right-censored. In this work, we study, for the first time, the entire trajectory of AD progression, which spans transitions from cognitively normal to MCI to AD. We model this disease trajectory as a multi-state process under intermittent observation, and formulate the effects of imaging and clinical covariates via functional mixed proportional intensity models, with imaging data as a functional covariate. We integrate functional principal component analysis (FPCA) with nonparametric maximum likelihood estimation and develop a stable EM algorithm for efficient computation. Moreover, we propose a profile score test to assess the association between the functional covariate and the multi-state process. We establish the asymptotic properties of the proposed estimators and test statistic through novel use of theories of FPCA, empirical processes, and semiparametric efficiency. Simulation studies and application to the Alzheimer's Disease Neuroimaging Initiative demonstrate the satisfactory performance of our proposed methods.

Yangjianchen Xu, University of Waterloo, Canada

Robust inference for the Cox proportional hazards model with interval-censored data

We establish the asymptotic properties of the nonparametric maximum likelihood estimators for potentially misspecified Cox proportional hazards models with general interval-censored data. Under mild regularity conditions, we show that the estimators for the regression parameters and the cumulative baseline hazard function converge almost surely to the minimizer of the Kullback–Leibler distance between the posited and true models. In addition, the estimators for the regression parameters are asymptotically normal with a limiting covariance matrix that can be consistently estimated by a sandwich estimator. Based on these asymptotic properties, we derive robust Wald and score statistics that can be used to perform valid statistical inference under various forms of model misspecification. Two important examples of such robust inference procedures are testing the treatment effect in a randomized clinical trial and identifying the ratios of the covariates' effects when the covariates satisfy certain conditions. Finally, we demonstrate the usefulness of the proposed robust inference procedures through extensive simulation studies and analysis of the UK Biobank data.

Kin Yau (Alex) Wong, The Hong Kong Polytechnic University, Hong Kong

A unified two-step estimation approach for semiparametric models under two-phase sampling

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 propose a novel approach for fitting semiparametric models to two-phase study data. We develop a general two-step estimation method that refines a complete-data estimator by incorporating the incomplete data, and auxiliary information if available, to improve efficiency. This unified framework performs estimation and inference on both Euclidean and infinite-dimensional parameters. One key advantage of the method is that it does not require modeling of the distribution of the missing covariates, and the resulting estimator is guaranteed to be at least as efficient as the complete-data estimator. We demonstrate the versatility of this approach through applications to transformation models for censored survival data and regression models with nonlinear effects. We establish theoretical properties of the proposed method, including estimation consistency and inference validity. We evaluate the performance of the proposed method via simulation studies and provide an application to a major cancer study.

IP49 — Stochastic Processes and Statistics

June 16 (Tuesday) 09:00–10:40 **Room:** 202 Program

Masahiro Kurisaki, RIKEN, Japan

Asymptotic Expansion of Nonlinear Filtering

The filtering problem is the problem of estimating an unobservable state in a time-series model, and it naturally arises in many areas such as control theory, weather forecasting, and finance. In this talk, we propose a novel asymptotic expansion approach that analytically expands the conditional distribution with respect to the intensity of the system noise. We show that the computation of the expansion coefficients reduces to solving stochastic (ordinary) differential equations incorporating the observations, which can be viewed as an extension of the Kalman–Bucy filter. Consequently, the expansion can be computed at a much lower cost than stochastic partial differential equation based methods, without relying on linear or Gaussian approximations as in the extended Kalman filter or the ensemble Kalman filter.

Hirofumi Shiba, The Institute of Statistical Mathematics, Japan

Diffusive Scaling Limits & Early Diagnostics for Piecewise Deterministic Monte Carlo Samplers

Piecewise deterministic Markov process (PDMP) samplers are attractive alternatives to Metropolis–Hastings methods, but their performance hinges on how partial velocity refreshment is implemented. We develop a high-dimensional scaling analysis to compare Forward Event-Chain Monte Carlo (FECMC) with the Bouncy Particle Sampler (BPS). For a standard Gaussian target, the (rescaled) negative log-density process converges to an Ornstein–Uhlenbeck limit with a diffusion coefficient strictly larger for FECMC than for optimally tuned BPS. The analysis implies that FECMC is asymptotically most efficient without global refreshment, and experiments show an approximately 15-fold gain in ESS per unit time. We also discuss an interesting application of our results to PDMP output analysis and diagnostics.

Teppei Ogihara, The University of Tokyo, Japan

Asymptotically uniformly most powerful tests for diffusion processes with nonsynchronous observations

We introduce a quasi-likelihood ratio testing procedure for diffusion processes observed under nonsynchronous sampling schemes. High-frequency data, particularly in financial econometrics, are often recorded at irregular time points, challenging conventional synchronous methods for parameter estimation and hypothesis testing. To address these challenges, we develop a quasi-likelihood framework that accommodates irregular sampling while integrating adaptive estimation techniques for both drift and diffusion coefficients, thereby enhancing optimization stability and reducing computational burden. We rigorously derive the asymptotic properties of the proposed test statistic, showing that it converges to a chi-squared distribution under the null hypothesis and exhibits consistency under alternatives. Moreover, we establish that the resulting tests are asymptotically uniformly most powerful. Extensive numerical experiments corroborate the theoretical findings and demonstrate that our method outperforms existing nonparametric approaches. This is a joint work with Futo Ueno.

Shogo Nakakita, The University of Osaka, Japan

Dimension-free uniform concentration bound for logistic regression

We study a dimension-free uniform concentration bound for logistic regression, a fundamental linear classification problem. While uniform concentration bounds are classical and fundamental objects in empirical process theory, deriving dimension-independent bounds for concrete statistical models has attracted considerable attention over the past decade. In this work, we establish a uniform concentration bound for logistic regression over Euclidean balls. In particular, we obtain a bound that yields a more natural sufficient condition for a uniform law of large numbers than those obtained via a naive application of Rademacher complexity arguments combined with McDiarmid's inequality. Our proof is based on a PAC-Bayes argument, combined with perturbations by Brownian motion and a second-order expansion via Itô's formula.

IP32 — Recent Advances in Theories and Methodologies for High-dimensional PCA

June 16 (Tuesday) 09:00–10:40 **Room:** 214 Program

Koji Tsukuda, Kyushu University, Japan

Multivariate allometric regression: methods and theory

The allometric extension model and its variants are useful for understanding growth and scaling in multivariate systems. We propose a new estimation framework for multivariate allometric regression in which the expected change in the response vector across values of the explanatory variables is constrained to lie in the first principal component direction of the response covariance matrix. This study focuses on estimating the leading principal component direction within the multivariate allometric regression model. We introduce a class of estimators encompassing conventional methods and establish sufficient conditions for consistency across multiple asymptotic regimes. We also present results from numerical simulations.

Kazuyoshi Yata, University of Tsukuba, Japan

Asymptotic Properties of Automatic Sparse PCA for High-Dimensional Data and Its Applications

High-dimensional data often exhibit a low-rank structure characterized by strongly spiked eigenvalues. In this talk, we investigate the estimation of such eigenstructures. Sparse principal component analysis (SPCA) has been extensively studied as a framework for estimating sparse principal component directions in high-dimensional settings. Aoshima and Yata (Statistica Sinica, 35) recently proposed Automatic SPCA (A-SPCA), which adaptively determines its threshold and has been shown to achieve consistency under mild conditions. We first show that A-SPCA can effectively estimate the underlying low-rank structure and establish its asymptotic properties under strongly spiked eigenvalue models. Furthermore, we show that this approach enables accurate estimation not only of the low-rank structure but also of the entire covariance matrix. In addition, we demonstrate that applying the proposed technique to kernel PCA allows for precise identification of nonlinear low-rank structures and yields high clustering accuracy. Applications to gene expression data further confirm the effectiveness of the proposed methods in practice. This work is joint research with Professor Makoto Aoshima (University of Tsukuba).

Shao-Hsuan Wang, National Central University, Taiwan

Bayesian sparse principal coordinates analysis with microbiome discoveries

Principal coordinates analysis (PCoA) is a foundational tool for exploring relationships among samples using dissimilarity information. However, classical PCoA produces dense loading vectors, making interpretation difficult—especially in ultrahigh-dimensional microbiome studies where pinpointing biologically relevant features is crucial. In this talk, I will introduce a sparse PCoA framework that incorporates regularization to yield interpretable, feature-selected coordinate axes. To flexibly accommodate a wide range of sparsity-inducing structures, we further develop a Bayesian sparse PCoA model that employs global–local shrinkage priors from the three-parameter beta–normal family to achieve adaptive sparsity. I will also present asymptotic results that clarify how sample size and dimensionality affect PCoA behavior. Through simulations and analyses of the Hadza gut microbiome and pancreatic cancer tumor microbiome datasets, we show that the proposed method substantially improves sparsity and interpretability while preserving key ecological and clinical signals.

CS02 — Causal Inference and Treatment Effects

June 16 (Tuesday) 09:00–10:40 **Room:** 215 Program

Taehyeon Koo, Columbia University

Distributionally Robust Synthetic Control: Ensuring Robustness Against Highly Correlated Controls and Weight Shifts

The synthetic control method estimates the causal effect by comparing the treated unit's outcomes to a weighted average of control units that closely match its pre-treatment outcomes, assuming the relationship between treated and control potential outcomes remains stable before and after treatment. However, the estimator may become unreliable when these

relationships shift or when control units are highly correlated. To address these challenges, we introduce the Distributionally Robust Synthetic Control (DRoSC) method, which accommodates potential shifts in relationships and addresses high correlations among control units. The DRoSC method targets a novel causal estimand defined as the optimizer of a worst-case optimization problem considering all possible weights compatible with the pre-treatment period. When the identification conditions for the classical synthetic control method hold, the DRoSC method targets the same causal effect as the synthetic control; when these conditions are violated, we demonstrate that this new causal estimand is a conservative proxy for the non-identifiable causal effect. We further show that the DRoSC estimator's limiting distribution is non-normal and propose a novel inferential approach. We demonstrate its performance through numerical studies and an analysis of the economic impact of terrorism in the Basque Country.

Yuming Sun, William & Mary

Estimating Heterogeneous Treatment Effects with Survival Outcomes via Deep Survival Learner

Understanding heterogeneity in treatment effects is central to precision medicine, yet its estimation remains challenging in survival settings with right censoring and time-varying effects. The conditional average treatment effect (CATE) provides a natural framework for characterizing individual-level treatment heterogeneity, but most existing methods for survival outcomes focus on a single prespecified time point and fail to exploit dependence across time, leading to inefficient or unstable estimation. We propose a Deep Survival Learner (DSL) for estimating heterogeneous treatment effects in right-censored survival settings. The method is built on a novel doubly robust pseudo-outcome whose conditional expectation identifies the CATE at each time point, ensuring robustness to misspecification of either the outcome model or the treatment assignment mechanism, provided censoring is correctly handled. To estimate CATEs over a clinically relevant time interval, DSL employs a multi-output deep neural network with shared hidden layers, enabling joint learning of treatment effects across time while borrowing strength from temporal dependence. Cross-fitting is used throughout to mitigate overfitting and bias from nuisance estimation. We evaluate the finite-sample performance of DSL through extensive simulations under multiple nuisance model misspecification scenarios. An application to the Boston Lung Cancer Study shows substantial heterogeneity in the effects of perioperative chemotherapy across patient characteristics and over time, demonstrating the practical value of the proposed approach.

Richard Guo, University of Michigan

Hunt-and-test strategies for ML-powered hypothesis testing

We consider designing specification or significance tests for nonparametric and semiparametric models, such as testing the goodness of fit of generalised additive models and detecting the heterogeneity of treatment effects. We propose a "hunt-and-test" strategy that involves splitting the data into two parts. On one part, we hunt for any signal that may be present in the score-type residuals following a fit of the null model. On the remaining data, we test for the presence of the potential signal thus found. For hunting, our framework allows the practitioner to use any flexible ML algorithm, such as a random forest, to detect complex alternatives. For calibrating the test, the first-order bias in the residuals is an obstacle that may lead to rejection under the null. To address this, we employ a debiasing step, which we show is equivalent to performing a certain weighted least squares regression. We establish that the type-I error can be controlled under relatively mild conditions and that the test has power against alternatives whenever the hunted signal is correlated with the true signal. Additionally, we show that by performing the data splitting multiple times and properly aggregating the tests, the procedure is stabilized and achieves even higher power.

Ha-Young Shin, Soongsil University

Treatment Effects on Hadamard Spaces

We propose a new framework for causal inference for Hadamard space-valued response variables, in which treatment effects are defined by a non-negative real number representing magnitude and a point at infinity representing direction. In particular, the average treatment effect (ATE) is defined as a magnitude-direction pair satisfying the following: when the the control group is transported a distance equal to that magnitude in that direction, its Frechet mean coincides with that of the treatment group. We define an ATE estimator explore its properties, including strong consistency, and detail an algorithm for computation of the ATE using Broyden's method. Simulations in which we investigate consistency and

confidence region coverage are included. We perform causal inference on real diffusion tensor imaging (DTI) data and compare with Euclidean alternatives. The interpretability of our ATE provides a major advantage over existing attempts to generalize the ATE to non-linear spaces; this is also illustrated with the DTI data.

Hisayuki Hara, Kyoto University

Spatial Statistical Models for Obsidian Source Composition

Understanding the distribution patterns of obsidian during the Jomon period in Japan (ca. 13,000 BC to 300 BC) is crucial for elucidating prehistoric human behavior. Because obsidian exhibits distinct chemical compositions depending on its geological source, it is possible to identify the provenance of excavated artifacts. The distribution of source composition ratios has therefore been widely used in archaeological studies to trace the movement of materials. Conventional approaches have primarily relied on descriptive statistics that aggregate source composition ratios across sites. However, several challenges remain, including estimating patterns in unsurveyed regions, modeling spatial dependence among nearby sites, accounting for differences in sample sizes across sites, and quantifying uncertainty in the estimates.

In this study, we propose a Multinomial Marked Cox Process (MMCP) as an integrated Bayesian spatial statistical model for estimating the spatial distribution of source composition ratios using presence-only data on obsidian finds, where only site locations are recorded, and absence information is unavailable. The MMCP accommodates presence-only data through a pseudo-absence point process and models spatial dependence using a Nearest-Neighbor Gaussian Process (NNGP) within a Bayesian multinomial logit framework. Efficient Gibbs sampling is achieved via Pólya–Gamma data augmentation. Furthermore, we incorporate prior distributions based on distance from each source, reflecting the archaeological assumption that the probability of obsidian from a given source decreases with increasing distance.

We apply the proposed method to obsidian data from the Kanto region of Japan and estimate the spatial distribution of source composition ratios across five phases, from the Early to the Late Jomon period. The results are presented and discussed from an archaeological perspective.

This is a joint work with Aru Ohta (Kyoto University) and Hiroomi Tsumura (Doshisha University).

DL15 — Frontiers in Non-Euclidean Data Analysis

June 16 (Tuesday) 11:00–12:40 **Room:** LT1A Program

Sungkyu Jung, Seoul National University, South Korea

Generalized Frechet Means and their applications

In this talk, I introduce generalized Fréchet means, a novel extension of classical Fréchet means that provides a unified framework for characterizing random elements in general metric spaces. The Fréchet mean extends the notion of an arithmetic average from Euclidean spaces to arbitrary metric spaces by defining it as the minimizer of the expected squared distance to the data. We generalize the Fréchet mean in several fundamental directions: by replacing the squared loss with an arbitrary cost function; by allowing the data space and parameter space to be distinct; and, most importantly, by permitting the domain of minimization for the empirical generalized Fréchet mean to be random and potentially different from that of its population counterpart. This added flexibility substantially broadens the scope of the Fréchet mean framework and enables its application to a wide range of statistical problems, including sequential dimension reduction for non-Euclidean data. We establish a strong consistency theorem for generalized Fréchet means and demonstrate the practical utility of the proposed framework by proving the consistency of principal geodesic analysis on the hypersphere, as well as an asymptotic guarantee for k-medoids clustering.

Zhenhua Lin, National University of Singapore, Singapore

Fréchet Single-Index Regression: Regularization, Estimation, and Optimality

Single-index models play an important role in semiparametric regression and connect naturally to ideas in dimension reduction and modern machine learning. Recent work has extended these models to Fréchet single-index regression, allowing the response to lie in a non-Euclidean metric space, such as covariance matrices, network Laplacians, or probability distributions. In this talk, I will present a regularized Fréchet single-index framework for analyzing such metric-valued responses.

The approach combines a total-variation–based regularization of the link function with an efficient estimation procedure for the index coefficients. I will summarize the main theoretical results, such as optimal convergence for the nonparametric component and root-n consistency for the index estimator, and illustrate the method's performance on both simulated and real datasets.

Jongmin Lee, Pusan National University, Korea

Huber means on Riemannian manifolds

This article introduces Huber means on Riemannian manifolds, providing a robust alternative to the Fréchet mean by integrating elements of both quadratic and absolute loss functions. The Huber means are designed to be highly resistant to outliers while maintaining efficiency, making it a valuable generalization of Huber's M-estimator for manifold-valued data. We comprehensively investigate the statistical and computational aspects of Huber means, demonstrating their utility in manifold-valued data analysis. Specifically, we establish nearly minimal conditions for ensuring the existence and uniqueness of the Huber mean and discuss regularity conditions for unbiasedness. The Huber means are consistent and enjoy the central limit theorem. Additionally, we propose a novel moment-based estimator for the limiting covariance matrix, which is used to construct a robust one-sample location test procedure and an approximate confidence region for location parameters. The Huber mean is shown to be highly robust and efficient in the presence of outliers or under heavy-tailed distributions. Specifically, it achieves a breakdown point of at least 0.5, the highest among all isometric equivariant estimators, and is more efficient than the Fréchet mean under heavy-tailed distributions. Numerical examples on spheres and the space of symmetric positive-definite matrices further illustrate the efficiency and reliability of the proposed Huber means on Riemannian manifolds. The proposed Huber mean can be implemented via 'geomstat' Python library.

DL19 — On Sinkhorn Semigroups and Related Fields

June 16 (Tuesday) 11:00–12:40 **Room:** LT1B Program

Pierre Del Moral, INRIA, France & The Chinese University of Hong Kong-Shenzhen, China

New Trends in the Stability Analysis of Sinkhorn Semigroups

Entropic optimal transport has become a central tool in machine learning and generative modeling, where Schrödinger bridges provide a scalable alternative to classical transport maps via Sinkhorn's algorithm. Understanding the stability of these iterative schemes as the number of iterations grows is a key challenge, traditionally addressed through nonlinear Perron–Frobenius theory, Hilbert projective metrics, and variational or PDE-based methods.

In this talk, we present a unified and self-contained perspective on recent advances in the stability analysis of Sinkhorn-type dynamics through semigroup methods. The approach relies on contraction coefficients and Lyapunov operator techniques, combined with tools from transportation inequalities (including log-Sobolev and Talagrand inequalities), φ-divergences, and Wasserstein geometry. These methods provide a flexible and powerful framework that both simplifies existing proofs and extends them to more general settings.

We highlight how this semigroup viewpoint yields new contraction estimates in terms of generalized entropy functionals, weighted total variation norms, and Kantorovich-type criteria, offering a coherent framework for analyzing the long-time behavior and robustness of Sinkhorn algorithms.

Xin Tong, National University of Singapore, Singapore

Wasserstein gradient flow for optimal distribution decomposition

We examine the infinite-dimensional optimization problem of finding a decomposition of a probability measure into K probability sub-measures to minimize specific loss functions inspired by applications in clustering and user grouping. We analyze the structures of the support of optimal sub-measures and introduce algorithms based on Wasserstein gradient flow, demonstrating their convergence. Numerical results illustrate the implementability of our algorithms and provide further insights.

Ajay Jasra, The Chinese University of Hong Kong-Shenzhen, China

New Trends in the Stability of Sinkhorn Semigroups

Entropic optimal transport problems play an increasingly important role in machine learning and generative modelling. In contrast with optimal transport maps which often have limited applicability in high dimensions, Schrodinger bridges can be solved using the celebrated Sinkhorn's algorithm, a.k.a. the iterative proportional fitting procedure. The stability properties of Sinkhorn bridges when the number of iterations tends to infinity is a very active research area in applied probability and machine learning. Traditional proofs of convergence are mainly based on nonlinear versions of Perron-Frobenius theory and related Hilbert projective metric techniques, gradient descent, Bregman divergence techniques and Hamilton-Jacobi-Bellman equations, including propagation of convexity profiles based on coupling diffusions by reflection methods. The objective of this review article is to present, in a self-contained manner, recently developed Sinkhorn/Gibbs-type semigroup analysis based upon contraction coefficients and Lyapunov-type operator-theoretic techniques.

IP53 — Machine Learning-informed Inference and Decision Making

June 16 (Tuesday) 11:00–12:40 **Room:** 209A Program

Yichen Zhang, Purdue University, United States

A Sparse Learning Framework for the High-Dimensional Newsvendor

Firms leveraging high-dimensional data for the feature-based newsvendor problem face a dual challenge: the operational risk of learning ineffective policies from a vast feature space and the strategic risk of leaking sensitive data. Standard privacy mechanisms are ill-suited for this context, as the noise required for privacy often scales with the ambient dimension, overwhelming the sparse signals that drive demand. To resolve this tension, we develop a differentially private algorithm that synergistically combines a noisy variant of Iterative Hard Thresholding with convolution-based smoothing of the newsvendor cost function. By carefully calibrating noise within the hard-thresholding step, our method simultaneously enforces model sparsity and provides rigorous privacy guarantees, while the smoothed surrogate stabilizes the iterative optimization process. We establish non-asymptotic error bounds that characterize the trade-offs among privacy, sparsity, feature dimension, and sample size. Crucially, our analysis reveals that the algorithm's performance depends on the intrinsic sparsity (s) rather than the ambient dimension (p). We prove that the excess operational cost (regret) converges at an accelerated rate of $O(sn^{-1}\log p)$, a significant improvement over the standard rate O(p^{1/2}$n^{-1/2}$) for non-smooth problems. These theoretical insights provide a principled justification for aggressive feature engineering, enabling firms to leverage complex data sources without incurring a dimensionality penalty. An empirical study using real-world retail data corroborates our theory, demonstrating that the proposed framework provides a robust solution for achieving both operational efficiency and data security in modern inventory management.

Sriram Sankararaman, University of California, Los Angeles, United States

Recent advances and challenges in imputation methods for large-scale biomedical data

Missingness is a pervasive problem in real-world biomedical datasets with the potential to adversely affect downstream inferential tasks. While many techniques to estimate or impute missing data have been proposed, missing data imputation in biomedical datasets remains a challenging problem due, in part, to the complex mechanisms that lead to missingness.

I will describe new approaches to imputation that are statistically expressive and scalable to large-scale biomedical datasets while being effective in settings where missingness is structured. I will then present results showing that these methods improve imputation accuracy across diverse settings including when applied to collections of traits or phenotypes measured across nearly 300,000 individuals from the UK Biobank. Further, the imputed phenotypes lead to a substantial increase in power to make genetic discoveries demonstrating the utility of our approach.

Heng Lian, City University of Hong Kong, Hong Kong

Distributed semi-supervised inference for generalized linear models with block-wise missing covariates

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.

Qiyang Han, Rutgers University, United States

Bandit algorithms: Precise dynamics and statistical inference

The multi-armed bandit problem is one of the most fundamental models in modern sequential decision-making. Although a large literature has focused on developing regret-optimal bandit algorithms, regret optimality alone does not guarantee valid statistical inference.

This talk takes a different perspective by studying the statistical inferential capacity of bandit algorithms through their precise arm-pull dynamics. We investigate this question for two canonical algorithms: Upper Confidence Bound (UCB) and Thompson sampling. Although both are known to be nearly regret-optimal, we show that their arm-pull behaviors are qualitatively different. In particular, for UCB all arms are stable with asymptotically deterministic arm-pull counts, whereas for Thompson sampling, such stability holds only for suboptimal arms and the unique optimal arm. This dichotomy reveals a unifying principle behind many existing (in)stability results for bandit algorithms: an arm is stable if and only if its interaction with statistical noise is asymptotically negligible. As a consequence, Wald-type confidence intervals are valid for all arms under UCB, but fail for optimal arms with multiplicity at least two under Thompson sampling.

Time permitting, we will also discuss the new technical approaches used to characterize these precise arm-pull dynamics, which are of a rather different nature for the two algorithms.

This talk is partially based on joint with Koulik Khamaru and Cun-Hui Zhang.

IP64 — Advances in Statistical and Machine Learning Methods for Biomedical Applications

June 16 (Tuesday) 11:00–12:40 **Room:** 209B Program

Sehwan Kim, Ewha Womans University, South Korea

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

In survival analysis, estimating the conditional survival function given predictors is often of interest. There is a growing trend in the development of deep learning methods

for analyzing censored time-to-event data, especially when dealing with high-dimensional predictors that are complexly interrelated. Many existing deep learning approaches for estimating the conditional survival functions extend the Cox regression models by replacing the linear function of predictor effects by a shallow feed-forward neural network while maintaining the proportional hazards assumption. Their implementation can be computationally intensive due to the use of the full dataset at each iteration because the use of batch data may distort the at-risk set of the partial likelihood function. To overcome these limitations, we propose a novel deep learning approach to non-parametric estimation of the conditional survival functions using the generative adversarial networks leveraging self-consistent equations. The proposed method is model-free and does not require any parametric assumptions on the structure of the conditional survival function. We establish the convergence rate of our proposed estimator of the conditional survival function. In addition, we evaluate the performance of the proposed method through simulation studies and demonstrate its application on a real-world dataset.

Taehwa Choi, Sungshin Women's University, South Korea

Interval-censored linear quantile regression

Censored quantile regression has emerged as a prominent alternative to classical Cox's proportional hazards model or accelerated failure time model in both theoretical and applied statistics. While quantile regression has been extensively studied for right-censored survival data, methodologies for analyzing interval-censored data remain limited in the survival analysis literature. This paper introduces a novel local weighting approach for estimating linear censored quantile regression, specifically tailored to handle diverse forms of interval-censored survival data. The estimation equation and the corresponding convex objective function for the regression parameter can be constructed as a weighted average of quantile loss contributions at two interval endpoints. The weighting components are nonparametrically estimated using local kernel smoothing or ensemble machine learning techniques. To estimate the nonparametric distribution mass for interval-censored data, a modified EM algorithm for nonparametric maximum likelihood estimation is employed by introducing subject-specific latent Poisson variables. Empirical performance of the proposed method is demonstrated through extensive simulation studies and real data analyses of two HIV / AIDS datasets.

Guanxun Li, Beijing Normal University at Zhuhai, China

E-value Aggregation via Data-Dependent Weighting and Its Application to Omics-Wide Differential Analysis

Motivated by recent findings establishing an equivalence between specific p-value-based multiple testing procedures and the e-Benjamini-Hochberg procedure, we develop a general framework for constructing novel multiple testing methods via the aggregation and combination of e-values. Since direct aggregation or combination often yields negligible power in practice, we introduce a data-dependent weighting scheme that significantly enhances the power of the resulting e-BH procedures. The design of these weights is nontrivial and draws inspiration from leave-one-out analysis, a technique widely employed to demonstrate false discovery rate control in p-value-based methods. We provide theoretical guarantees that the proposed e-Benjamini-Hochberg procedure, when equipped with these data-dependent weights, achieves finite-sample FDR control. Building upon this weighting framework, we propose new procedures tailored for Omics-Wide Differential Analysis: (i) a method for assembling e-values derived from distinct data subsets, ensuring simultaneous control of both group-wise and overall FDRs; and (ii) adaptive multiple testing methods that leverage external structural information to boost power. Numerical studies demonstrate the effectiveness and advantages of the proposed methods across these application scenarios.

Junsouk Choi, Korea University, South Korea

Semi-supervised Spatial Topic Modeling for Discovery of Multicellular Spatial Tissue Structures in Multiplex Imaging

Deciphering spatial architecture of tissues is crucial for understanding complex cellular interactions and their implications for disease pathology and clinical outcomes. Recent advances in multiplexed imaging now enable high-resolution profiling of cell phenotypes and their spatial arrangements, revealing the pivotal role of tissue structures in modulating immune responses and driving disease progression. To systematically identify and characterize such structures, we propose a novel semi-supervised Bayesian spatial topic model that integrates spatial Gaussian processes into latent Dirichlet allocation to flexibly capture spatial dependencies inherent in tissue organization. By jointly modeling multiple multiplexed images, the proposed approach identifies consistent and coherent spatial structures across samples and incorporates clinical covariates to further guide and refine these discoveries. Applied to a lung cancer multiplexed imaging dataset, our approach reveals biologically meaningful tumor microenvironment patterns that are consistent across patients and significantly associated with clinical outcomes.

IP68 — Statistical and Optimization Perspectives of Generative Models

June 16 (Tuesday) 11:00–12:40 **Room:** 203 Program

Masashi Sugiyama, The University of Tokyo, Japan

Recent Advances in Reward Modeling for Reinforcement Learning

Reinforcement learning (RL) has achieved remarkable success in robotics, games, and language model post-training, where

reward signals are essential for effective agent training. In this talk, I will present our recent research on advanced reward modeling. This includes robustifying RL through transfer and weakly supervised learning, coping with interval-based rewards, and exploring diverse reward aggregation frameworks that move beyond the traditional discounted sum.

Xiuyuan Cheng, Duke University, United States

Minimax learning in Wasserstein space via neural transport maps

We study minimax learning problems arising in distributionally robust optimization (DRO), where an adversary perturbs the data distribution within Wasserstein space. By parameterizing distributions as pushforwards of a reference measure via neural transport maps, we reduce the infinite-dimensional problem to optimization over functions. This enables a practical gradient descent-ascent (GDA) scheme that alternates between updating the model and the transport map, with global convergence guarantees under conditions beyond convex–concave settings. The learned transport map also provides a direct way to generate from adversarial distributions without costly inner optimization.

Yuting Wei, University of Pennsylvania, United States

Dimension-Free Convergence of Diffusion Models for Approximate Gaussian Mixtures

Diffusion models are distinguished by their exceptional generative performance, particularly in producing high-quality samples through iterative denoising. While current theory suggests that the number of denoising steps required for accurate sample generation should scale linearly with data dimension, this does not reflect the practical efficiency of widely used algorithms like Denoising Diffusion Probabilistic Models (DDPMs). This paper investigates the effectiveness of diffusion models in sampling from complex high-dimensional distributions that can be well-approximated by Gaussian Mixture Models (GMMs). For these distributions, our main result shows that DDPM takes at most O(1/eps) iterations to attain an eps-accurate distribution in total variation (TV) distance, independent of both the ambient dimension d and the number of components K, up to logarithmic factors. Furthermore, this result remains robust to score estimation errors. These findings highlight the remarkable effectiveness of diffusion models in high-dimensional settings given the universal approximation capability of GMMs, and provide theoretical insights into their practical success.

Yuan Yao, The Hong Kong University Science and Technology, Hong Kong

TBD

TBD

IP71 — IMS New Researchers Invited Session

June 16 (Tuesday) 11:00–12:40 **Room:** 201 Program

Xiaozhu Zhang, University of Washington, United States

Convex Mixed-Integer Programming for Causal Additive Models with Optimization and Statistical Guarantees

Learning causal structure from nonlinear data remains computationally and statistically challenging. We address this problem for additive, nonlinear structural equation models with Gaussian noise. By expressing each non-linear function through a basis expansion, we formulate a maximum likelihood estimator with a group l0 penalty that directly controls graph sparsity. Despite the combinatorial nature of the problem, the estimator can be solved efficiently via a convex mixed-integer program combined with a novel outer approximation scheme, yielding globally optimal solutions within seconds. Our framework can incorporate prior knowledge such as edge constraints and partial orders. We prove consistency for graph recovery when the number of variables grows with sample size, and connect optimization guarantees to statistical error, yielding an early stopping rule that preserves consistency while reducing computation. Compared to methods based on linearity, equal noise variances, or heuristics, our approach provides a unified framework with provable optimality, efficiency,

and statistical guarantees. Experiments demonstrate substantial improvements in graph recovery on both simulated and real data. This is joint work with Nir Keret, Ali Shojaie, and Armeen Taeb.

Charlie Wolock, University of Rochester, United States

Leveraging machine learning to estimate survival curves with current status data

In many epidemiological study designs, time-to-event outcomes may be subject to current status sampling: rather than observing the outcome itself, the investigator observes each study participant at a single monitoring time, recording a binary indicator of whether the event has occurred by that time. Such study design results in an extreme form of interval censoring. Existing nonparametric methods for current status data typically require independence between the monitoring time and the event time, which may be unrealistic in practice. We propose an approach to estimating the survival curve of a time-to-event outcome under current status sampling using tools from semiparametric efficiency theory and shape-constrained estimation. Our method allows for monitoring processes that are informed by measured covariates and employs machine learning tools to flexibly estimate nuisance parameters. We devise a sensitivity analysis approach investigating the degree to which the resulting estimates change under deviations from conditionally uninformative monitoring. We use the proposed methods to estimate the duration of COVID-19 symptoms in a university population.

Julien Laurendeau, École Polytechnique Fédérale de Lausanne, Switzerland

New guarantees for optimal regimes in the presence of unmeasured confounding

We develop a framework for treatment regimes that improve on the expected value of conventional optimal regimes when there is effect-modifying unmeasured confounding. We first examine superoptimal regimes, decision functions that incorporate an individual's natural treatment value, the treatment that decision makers would assign in the absence of intervention. By construction, these regimes dominate classical optimal regimes and reduce to them when the standard no-unmeasured-confounding assumption is correct. When this assumption fails, superoptimal regimes can yield substantial value improvements. We show that their identification in observational settings requires no stronger assumptions than those used to identify conventional value functions, and we establish sharp bounds obtainable by straightforward transformations of existing bounds for optimal regimes. We then introduce initiation regimes in time-varying settings where human experts possess additional unrecorded information. These regimes allow human decisions to proceed until initiating the optimal dynamic regime improves expected outcomes. We develop experimental designs that identify the best initiation regimes and provide estimation and inference procedures To illustrate our the practical performance of our methods relative to conventional optimal regimes, we analyze the effect of prompt ICU admission on survival using instrumental-variable methods and the effect of dynamic treatment regimes on back pain treatment.

Jin-Hong Du, The University of Hong Kong, Hong Kong

Seeing Through Correlations: Disentangled Feature Importance

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 vulnerable to correlation distortion, limiting their robustness across diverse tasks. We introduce Disentangled Feature Importance (DFI), a model-agnostic framework that resolves these limitations by combining principled statistical inference with computational flexibility. DFI leverages 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.

CS05 — Time Series, Stochastic Processes, and Financial Statistics

June 16 (Tuesday) 11:00–12:40 **Room:** 202 Program

Gehui Zhang, Southwest Petroleum University

Multivariate Stochastic Volatility with Informative Missingness

Multivariate stochastic volatility (MSV) models, which treat the time-varying covariance matrix of a multivariate time series as a stochastic process, are essential for characterizing dynamic variability and co-dependencies. Existing methods for MSV modeling are largely constrained by the assumption that data are missing at random. However, modern technologies increasingly generate high-dimensional, self-reported time series data in which missingness is inherently informative, limiting the applicability of current approaches. This article develops a novel statistical framework for MSV models with data that are missing not at random. We propose a multivariate imputation method based on a generalized Tukey's representation that leverages the joint Markovian structure of MSV models to mitigate unidentifiability in informative missingness settings. This imputation approach is integrated into a conditional particle filter with ancestral sampling, implemented within a particle Gibbs sampling scheme to account for imputation uncertainty. The proposed method's performance is demonstrated through simulation studies and an application to multivariate mobile phone self-reported mood data from an individual monitored following a suicide attempt.

Boyuan Ning, Waseda University

Estimation of the Elasticity for CKLS Model from High-Frequency Observations

We investigate parametric estimation of the elasticity parameter in the CKLS diffusion based on high-frequency data. First, we transform the CKLS diffusion to a CIR-type one via a smooth state-space mapping and the general Girsanov change of measure. This transformation enables the applications of existing inference tools for CIR processes while ensuring possibilities of transferring the resulting limit theorems back to the original probability space. However, because Feller's condition fails, many existing high-frequency likelihood-based procedures cannot be applied directly, since their discretization schemes approximate likelihood terms involving the reciprocal of the process by Riemann sums that are no longer well-defined once the paths are allowed to hit zero. Instead, we estimate the drift coefficient of the transformed CIR-type model via a procedure based on its positive Harris recurrence, which is valid in the high-frequency regime. Exploiting the drift-elasticity relationship implied by the CKLS–CIR transformation, with the help of an initial estimation, we obtain an estimator for the CKLS elasticity from the CIR drift estimator in the transformed model. This yields a closed-form estimator for the elasticity parameter with an explicit asymptotic variance. We establish its p-consistency, stable convergence in law, and asymptotic normality. Finally, we show that stable convergence in law is invariant under equivalent changes of measure, thereby guaranteeing that the Gaussian limit remains invariant under the original measure.

Edward Hill, Queen Mary University of London

Nonstationarity Extended Whittle Estimation of Cyclical Time Series

We develop a two-stage estimation procedure for time series containing a persistent cycle of unknown length. Such dynamics include stationary cyclical long memory as well as nonstationary and unit-root cycles, providing a unified framework for analysing persistent oscillatory behaviour. In the first stage, the length of the cycle is estimated by maximising the periodogram over the Fourier frequencies. The estimator is shown to be n-consistent and to possess a nonstandard limit distribution. This limit theory enables the construction of narrow confidence intervals and forms the basis for testing for the presence of a cycle. In the second stage, the memory parameter and the autoregressive and moving-average coefficients are estimated using a Whittle likelihood based on a corrected periodogram that incorporates the estimated frequency and accommodates nonstationarity. The resulting estimators are shown to be consistent and asymptotically normal under both stationarity and nonstationarity. Monte Carlo simulations indicate good finite sample performance of the two-stage procedure and empirical illustrations using influenza and exchange rate data are provided.

Jiazhen Xu, Macquarie University

Spherically Embedded Time Series with Unknown Trend and Periodic Components

Spherically embedded time series are time series with values naturally residing on or can be equivalently mapped to the unit sphere. Such data are increasingly prevalent in diverse fields with important examples being energy generation compositional time series and bike trip volume distributional time series. Despite their ubiquity, these data often exhibit complex non-stationarity driven by latent deterministic trend and periodic components. Traditional Euclidean time series methods fail to account for the intrinsic non-Euclidean geometry of the sphere, leaving a critical gap in rigorous methodologies for modeling and forecasting.

To address this methodological gap, we propose a unified geometric framework for the analysis of nonstationary spherically embedded time series. Central to our approach is a nonparametric spherical trend-periodicity decomposition model, which utilizes a optimal-transport-based removal operator to systematically decouple the deterministic components while preserving the spherical topology. The estimation procedure operates sequentially where we first estimate and extract the global smooth trend, followed by the identification of the unknown period and the extraction of the periodic structure. The resulting de-trended and de-seasonalized stationary residuals can be further modeled via a spherical autoregressive process. We formalize this semiparametric framework as the trend-periodic spherical autoregressive model. We establish the rigorous theoretical foundations for this modeling procedure, proving its consistency under temporal dependence. Extensive simulations corroborate these theoretical guarantees and demonstrate the superior finite-sample predictive performance of the trend-periodic spherical autoregressive model. Finally, we validate the practical utility of our methodology through applications to U.S. electricity generation compositions and bike trip volume profiles in New York City. Both empirical analyses yield significantly enhanced forecasting accuracy while providing interpretable, geometrically grounded insights into the underlying structural dynamics.

Ziyuan Zhang, Purdue University

Data Driven Asset Pricing

Classical asset pricing relies on the existence of an equivalent martingale measure (EMM), whose construction depends on specifying a probabilistic model for market dynamics. In practice, the true data-generating process is unknown, and model misspecification can lead to unreliable pricing and hedging.

We propose a data-driven, model-free asset pricing framework that constructs a martingale measure directly from observed data. Using a modified entropy minimization scheme combined with deep neural networks, our approach flexibly selects and approximates an EMM without assuming a parametric market model. The resulting method adapts to complex nonlinear structures in financial returns and provides stable pricing across diverse market environments.

CS06 — Survival Analysis, Clinical Trials, and Biostatistics

June 16 (Tuesday) 11:00–12:40 **Room:** 214 Program

HYEON SEOK Oh, Korea University

Identifiability in Semiparametric-Parametric Frailty Models with Dependent Competing Risks

TBD

Shanpeng Li, City of Hope

Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Competing Risks Time-to-Event Data: Applications to Mega-Scale Health Research

Joint modeling has become increasingly popular for the analysis of longitudinal and time-to-event data. However, semiparametric multivariate joint modeling for large-scale datasets still faces substantial statistical and computational challenges, primarily due to the high dimensionality of random effects and the complexity of estimating nonparametric baseline hazards. These issues often result in prolonged computation time and excessive memory usage, limiting the practical utility of joint modeling for biobank-scale analyses. In this article, we develop an approximate EM algorithm and an efficient implementation for a semiparametric multivariate joint modeling framework that substantially reduces computation

time and requires only moderate memory usage, enabling analyses involving large numbers of subjects and multiple longitudinal biomarkers. The scalability and estimation accuracy of our approach are demonstrated through simulation studies and an application to the UK Biobank primary care (UKB-PC) data, jointly modeling six longitudinal biomarkers and competing risks in over ten thousand subjects. We also illustrate that leveraging biobank-scale UKB data can improve dynamic prediction in both discrimination and calibration, while models trained on smaller samples may experience reduced accuracy, particularly for low-event or rare outcomes. A user-friendly R package, FastJM, has been developed and is publicly available on the Comprehensive R Archive Network: https://CRAN.R-project.org/package=FastJM.

Yanqing Yi, Memorial University of Newfoundland

Optimal Adaptive Randomized Clinical Trial with Covariates

This talk discusses the adaptive randomization for clinical trials using of patients' covariates to allocate more patients to the potential better treatments. Considering the average rewards, the allocation of treatments is formatted as a Markov decision process. The Bellman equation is established for the process, and the Bellman operator is proven to be a contractor for certain allocation rules. The optimal randomization is sequentially identified using the Bellman contraction operator. Numeric studies are conducted to evaluate the performance of the proposed method.

Abhimanyu Singh Yadav, Banaras Hindu University

On Generalized Adaptive Progressive First-Failure Censoring Scheme with Optimal Design

Life-testing experiments aim to identify censoring designs that preserve the efficiency of statistical estimators while reducing overall experimentation time, thereby saving cost and resources. This paper introduces a novel censoring scheme within a group-testing framework, which balances resource savings with estimator efficiency. The proposed design generalizes several existing censoring techniques, providing a flexible and practical approach for life-testing studies. Under this scheme, maximum likelihood estimators, asymptotic confidence intervals, and bias-corrected bootstrap confidence intervals are derived for the parameters of the Weibull distribution. Alternative estimation approaches, including the Expectation-Maximization (EM) and Stochastic EM algorithms, are also developed for censored samples. Methods for point prediction and prediction intervals are proposed to estimate the failure times of censored observations. A Monte Carlo simulation study evaluates the performance of the estimators and the expected experimentation time, comparing the proposed design with existing censoring schemes. Results show that the proposed scheme effectively reduces test time while maintaining or improving estimator efficiency. Optimal censoring designs are explored under various optimality criteria. Finally, a real-life engineering data set illustrates the applicability and advantages of the proposed censoring strategy, demonstrating its practical relevance for efficient life-testing experiments.

Yuji Komiyama, Tohoku University

A Finite-Time-Horizon Mixture Cure Model with Application to Online Marketplace Data

This study proposes a mixture cure model that latently divides a population based on event occurrence within a finite time horizon. Conventional models rely on event occurrence over an infinite horizon, introducing untestable assumptions that often lead to issues with identifiability and interpretability. By shifting the estimand to a specific period of interest, the proposed approach reduces reliance on these infinite-tail assumptions and aligns interpretations more closely with finite-horizon decision-making objectives. Through simulation studies, we first evaluate the statistical properties of the proposed estimator, including estimation bias and variance. We further show that relying on conventional infinite-horizon models for finite-horizon decision-making can lead to erroneous judgments. Finally, we apply the model to user behavior data from Mercari, a Japanese online flea market platform. The empirical results reveal that the proposed model identifies different significant variables compared to the conventional model, offering interpretations that better reflect seasonal variation in user behavior.

CS11 — Change Point Detection, High-Dimensional Time Series, and Functional Data

June 16 (Tuesday) 11:00–12:40 **Room:** 215 Program

Dylan Dijk, University of Bristol

Tail-Robust Change Point Detection in High-Dimensional Linear Regression with Non-Sparse Structures

Change point detection in high-dimensional linear regression has been shown to be feasible without assuming sparsity in either the regression parameters or their changes. However, existing results typically rely on sub-Gaussian or light-tailed assumptions, limiting their applicability in settings with heavy-tailed data. In this work, we develop a tail-robust methodology for multiple change point detection under weak moment conditions, requiring only finite $(2 + 2\epsilon)$-th moments for the covariates, and $(1 + \epsilon)$-th moments for the noise, for some $\epsilon \in (0, 1)$. Our approach is based on applying element-wise truncation to the product of the covariates and the response, which is then used to construct a change point statistic that measures discrepancies in their local covariance. We establish consistency of the proposed procedure under these minimal assumptions, thereby extending existing non-sparse change point detection results to heavy-tailed settings, with an estimation rate that depends on ϵ, hence quantifying the effect of tail heaviness. In addition, simulation studies demonstrate that the proposed method maintains strong statistical performance under heavy-tailed distributions compared to non-robust alternatives, while also preserving performance in the Gaussian case.

Debanjana Datta, Indian Statistical Institute

Detection of Structural Shifts in Functional Time Series

We have devised a novel procedure for change point detection in Functional Time Series (FTS) using a state space representation. Our model features a Functional Autoregressive (FAR) latent variable process driven by Gaussian innovations, naturally accommodating sparse, noisy observations. Crucially, unlike conventional methods, our approach avoids smoothing and Functional Principal Component Analysis (FPCA), thereby mitigating the substantial information loss typically incurred across the functional domain. An efficient Blocked Gibbs sampling algorithm is developed for estimating their locations. The methodology is efficient enough to detect regime shifts in the mean function, volatility and AR-operator. Furthermore, we shall illustrate examples having financial aspects.

Jaesung Park, Seoul National University

Principal Component Analysis for Zero-Inflated Compositional Data

Recent advances in DNA sequencing technology have led to a growing interest in microbiome data. Since the data are often high-dimensional, there is a clear need for dimensionality reduction. However, the compositional nature and zero-inflation of microbiome data present many challenges in developing new methodologies. New PCA methods for zero-inflated compositional data are presented, based on a novel framework called principal compositional subspace. These methods aim to identify both the principal compositional subspace and the corresponding principal scores that best approximate the given data, ensuring that their reconstruction remains within the compositional simplex. To this end, the constrained optimization problems are established and alternating minimization algorithms are provided to solve the problems. The theoretical properties of the principal compositional subspace, particularly focusing on its existence and consistency, are further investigated. Simulation studies have demonstrated that the methods achieve lower reconstruction errors than the existing log-ratio PCA in the presence of a linear pattern and have shown comparable performance in a curved pattern. The methods have been applied to four microbiome compositional datasets with excessive zeros, successfully recovering the underlying low-rank structure.

Namgil Lee, Kangwon National University

A Unified Framework for Shrinkage Estimation of High-Dimensional Vector Autoregressive Models with the R Package VARshrink

Shrinkage estimation is useful for high-dimensional vector autoregressive (VAR) models where the dimensionality exceeds the sample size. We present the R package VARshrink, which implements various nonparametric, parametric, and semi-parametric shrinkage estimation methods within a unified framework. For parametric and semi-parametric approaches, we implemented Bayesian VAR models with a scale mixture of multivariate normal distributions for the noise, enabling robust

estimation of correlations between noise variables in the presence of outliers. VAR estimation problems are expressed as multivariate regressions, making all methods accessible through the VARshrink() interface. In addition, the effective number of parameters can be derived from the shrinkage intensity parameter, allowing information criteria to be used for selecting the most suitable VAR model. Performance comparisons and real data application will be demonstrated.

Mingxu LI, Beijing Normal Hong Kong Baptist University

A GARCH-MIDAS-CJ Model: Integrating Jump Decomposition into Multiplicative Component Volatility

Standard GARCH-MIDAS models use realized volatility as a monolithic input to the long-term variance component, discarding the economically distinct information contained in its continuous and discontinuous parts. Once realized volatility is separated into a continuous component(C) and a jump component(J) using threshold-based estimators that are consistent and robust to finite-sample bias, the two components can be directed to where they naturally belong within the multiplicative variance structure. We propose a GARCH-MIDAS-CJ model in which the continuous component drives the long-term component, while the jump component enters the short-term equation. This specification embeds the distinction between persistent diffusive variation and transient price discontinuities directly into the conditional variance decomposition. Empirical analysis on S&P 500 high-frequency data shows that the jump sensitivity parameter is statistically significant, and that the proposed CJ decomposition improves both in-sample fit and out-of-sample forecast accuracy relative to the conventional GARCH-MIDAS specification driven by realized volatility alone. Results are robust to alternative jump detection methods and MIDAS weight specifications.

IP43 — Advances in Deep Learning and Kernel Learning Methods

June 16 (Tuesday) 13:30–15:10 **Room:** LT1A Program

Won Chang, Seoul National University, South Korea

Deep Normalizing Flow Methods for Fast Bayesian Inference

Bayesian methods based on Markov chain Monte Carlo (MCMC) enable straightforward statistical modeling and inference by sequentially sampling from posterior distributions. However, they often face computational challenges due to slow chain mixing, especially when the posterior is complex. Normalizing flows (NFs) provide a computationally efficient way to approximate posterior distributions. Equipped with modern deep neural networks for invertible transformations, NFs can fully exploit GPU-accelerated computing and thus overcome the fundamental limitations of MCMC arising from its sequential nature. In this talk, I introduce a few NF-based Bayesian inference methods that I have developed in recent years, which efficiently handle parameter boundaries, multimodality, and heavy-tailed posteriors.

Yuqi Gu, Columbia University, United States

Discrete Causal Representation Learning

Causal representation learning seeks to uncover causal relationships among high-level latent variables from low-level, entangled, and noisy observations. Existing approaches often either rely on deep neural networks, which lack interpretability and formal guarantees, or impose restrictive assumptions like linearity, continuous-only observations, and strong structural priors. These limitations particularly challenge applications with a large number of discrete latent variables and mixed-type observations. To address these challenges, we propose discrete causal representation learning (DCRL), a generative framework that models a directed acyclic graph among discrete latent variables, along with a sparse bipartite graph linking latent and observed layers. This design accommodates continuous, count, and binary responses through flexible measurement models while maintaining interpretability. Under mild conditions, we prove that both the bipartite measurement graph and the latent causal graph are identifiable from the observed data distribution alone. We further propose a three-stage estimate-resample-discovery pipeline: penalized estimation of the generative model parameters, resampling of latent configurations from the fitted model, and score-based causal discovery on the resampled latents. We establish the consistency

of this procedure, ensuring reliable recovery of the latent causal structure. Empirical studies on educational assessment and synthetic image data demonstrate that DCRL recovers sparse and interpretable latent causal structures.

Qian Lin, Tsinghua University, China

The adaptive feature program

We propose a framework to investigate some interesting phenomenon appeared in deep neural networks.

Feng Ruan, Northwestern University, United States

A Theory of Feature Learning in Kernel Models

We study feature learning in a compositional variant of kernel ridge regression in which the predictor is applied to a learnable linear transformation of the input. When the response depends on the input only through a low-dimensional predictive subspace, we show that all global minimizers of the population objective for the linear transformation annihilate directions orthogonal to this subspace, and in certain regimes, exactly identify the subspace. Moreover, we show that global minimizers of the finite-sample objective inherit the exact same low-dimensional structure with high probability, even without any explicit penalization on the linear transformation.

IP31 — Optimal Transport and Beyond

June 16 (Tuesday) 13:30–15:10 **Room:** LT1B Program

Ting-Kam Leonard Wong, University of Toronto, Canada

Shape constrained density estimation with Wasserstein projection

Statistical inference based on optimal transport offers a different perspective from that of maximum likelihood, and has increasingly gained attention in recent years. In this paper, we study univariate nonparametric shape-constrained density estimation via projection with respect to the p-Wasserstein distance, with a focus on the quadratic case $p = 2$. By considering shape constraints given by displacement convex subsets of the Wasserstein space, Wasserstein projection estimation is a convex optimization problem. We focus on two fundamental examples, namely non-increasing densities on the non-negative real line, and log-concave densities on the real line. In each case, we prove structural properties of the Wasserstein projection estimator, propose a discretization which can be implemented by off-the-shelf solvers, and compare the projection estimator with the corresponding maximum likelihood estimator.

Matthew Thorpe, University of Warwick, United Kingdom

Laplace Learning in Wasserstein Space

Laplace Learning is a graph-based method for semi-supervised learning. Given a set of partially labelled feature vectors the objective is to find the missing labels. Laplace Learning minimises a Dirichlet energy. Under suitable scaling conditions one can deduce a PDE limit as the number of feature vectors goes to infinity. Optimal transport plays a role in the discrete-to-continuum convergence of minimisers. Furthermore, we extend the analysis from finite dimensional Euclidean settings to a submanifold of the Wasserstein space.

Tin Lok James Ng, Trinity College Dublin, Ireland

Bayesian Spatially Varying Regression with Optimal Transport

We study spatial regression models in which the relationship between a response and covariates varies across a spatial domain. While existing approaches such as geographically weighted regression, Gaussian process based spatially varying coefficient models, and graph based clustering methods offer flexible modeling of spatial heterogeneity, they often suffer from sensitivity to tuning choices, reliance on arbitrary graph constructions, or limited ability to predict at unobserved

locations. We propose a new spatial partitioning framework based on optimal transport that directly partitions the entire spatial domain and supports out of sample prediction. Our approach uses Laguerre tessellations induced by semi discrete optimal transport to obtain flexible partitions. To address computational challenges, we also develop a fully discrete entropic regularized optimal transport formulation that yields efficient soft partitions defined on observed locations while still enabling prediction at new sites. Framed in a Bayesian setting, we place priors on target probability measures and spatially varying regression coefficients, and we study posterior contraction rates for both formulations. An efficient MCMC algorithm is developed for posterior inference in the fully discrete case, demonstrating the practical feasibility of the proposed methodology.

Yoshikazu Terada, The University of Osaka, Japan

A New Perspective on Matrix Decomposition Factor Analysis

Numerous estimators have been proposed for factor analysis, and their statistical properties have been extensively studied. In the early 2000s, a novel matrix factorization-based approach, known as Matrix Decomposition Factor Analysis (MDFA), was introduced and has been actively developed in computational statistics. The MDFA estimator offers several advantages, including the guarantee of proper solutions (i.e., no Heywood cases) and the extensibility to sparse estimation. However, the MDFA estimator does not appear to be formulated as a classical M-estimator or a minimum discrepancy function (MDF) estimator, and the statistical properties of the MDFA estimator have remained largely unexplored. Although the MDFA estimator minimizes a loss function resembling that of principal component analysis (PCA), it empirically behaves more like consistent estimators in factor analysis than like PCA itself. This raises a fundamental question: can matrix decomposition factor analysis truly be regarded as factor analysis"? To address this problem, we establish consistency and asymptotic normality of the MDFA estimator. By recognizing MDFA as a semiparametric maximum likelihood estimator, we show that its profile likelihood is given by the squared Bures–Wasserstein distance between the sample covariance matrix and the modeled covariance matrix. As a consequence, the MDFA estimator is ultimately an MDF estimator for factor analysis. Beyond MDFA, the same representation holds for a broad class of component analysis methods, including PCA, thereby offering a unified perspective on component analysis.

IP57 — Recent Advances in Variational Inference

June 16 (Tuesday) 13:30–15:10 **Room:** 209A Program

Kamélia Daudel, ESSEC Business School, France

Importance Weighted Variational Inference without the Reparameterization Trick

Importance weighted variational inference (VI) approximates densities known up to a normalizing constant by optimizing bounds that tighten with the number of Monte Carlo samples N. Standard optimization relies on reparameterized gradient estimators, which are well-studied theoretically yet restrict both the choice of the data-generating process and the variational approximation. While REINFORCE gradient estimators do not suffer from such restrictions, they lack rigorous theoretical justification. In this paper, we provide the first comprehensive analysis of REINFORCE gradient estimators in importance weighted VI, leveraging this theoretical foundation to diagnose and resolve fundamental deficiencies in current state-of-the-art estimators. Specifically, we introduce and examine a generalized family of variational inference for Monte Carlo objectives (VIMCO) gradient estimators. We prove that state-of-the-art VIMCO gradient estimators exhibit a vanishing signal-to-noise ratio (SNR) as N increases, which prevents effective optimization. To overcome this issue, we propose a novel VIMCO-type gradient estimator and show that it averts the SNR collapse of existing VIMCO gradient estimators. We demonstrate its superior empirical performance compared to current VIMCO implementations in challenging settings where reparameterized gradients are typically unavailable.

Dario Draca, The University of Sydney, Australia

Inversion-Free Natural Gradient Descent on Riemannian Manifolds

The natural gradient method is widely used in statistical optimization, but its standard formulation assumes a Euclidean parameter space. This paper proposes an inversion-free stochastic natural gradient method for probability distributions whose parameters lie on a Riemannian manifold. The manifold setting offers several advantages: one can implicitly enforce parameter constraints such as positive definiteness and orthogonality, ensure parameters are identifiable, or guarantee regularity properties of the objective like geodesic convexity. Building on an intrinsic formulation of the Fisher information matrix (FIM) on a manifold, our method maintains an online approximation of the inverse FIM, which is efficiently updated at quadratic cost using score vectors sampled at successive iterates. In the Riemannian setting, these score vectors belong to different tangent spaces and must be combined using transport operations. We prove almost-sure convergence rates of $O(\log s/s^{\alpha})$ for the squared distance to the minimizer when the step size exponent $\alpha > 2/3$. We also establish almost-sure rates for the approximate FIM, which now accumulates transport-based errors. A limited-memory variant of the algorithm with sub-quadratic storage complexity is proposed. Finally, we demonstrate the effectiveness of our method relative to its Euclidean counterparts on variational Bayes with Gaussian approximations and normalizing flows.

Cheng Zhang, Peking University, China

Provable Sample-Efficient Transfer Learning Conditional Diffusion Models via Representation Learning

While conditional diffusion models have achieved remarkable success in various applications, they require abundant data to train from scratch, which is often infeasible in practice. To address this issue, transfer learning has emerged as an essential paradigm in small data regimes. Despite its empirical success, the theoretical underpinnings of transfer learning conditional diffusion models remain unexplored. In this paper, we take the first step towards understanding the sample efficiency of transfer learning conditional diffusion models through the lens of representation learning. Inspired by practical training procedures, we assume that there exists a low-dimensional representation of conditions shared across all tasks. Our analysis shows that with a well-learned representation from source tasks, the sample complexity of target tasks can be reduced substantially. Numerical experiments are also conducted to verify our results.

Minh-Ngoc Tran, The University of Sydney, Australia

Bures-Wasserstein Importance-Weighted Evidence Lower Bound: Exposition and Applications

The Importance-Weighted Evidence Lower Bound (IW-ELBO) has emerged as an effective objective for variational inference (VI), tightening the standard ELBO and mitigating the mode-seeking behaviour. However, optimizing the IW-ELBO in Euclidean space is often inefficient, as its gradient estimators suffer from a vanishing signal-to-noise ratio (SNR). This paper formulates the optimisation of the IW-ELBO in Bures-Wasserstein space, a manifold of Gaussian distributions equipped with the 2-Wasserstein metric. We derive the Wasserstein gradient of the IW-ELBO and project it onto the Bures-Wasserstein space to yield a tractable algorithm for Gaussian VI. A pivotal contribution of our analysis concerns the stability of the gradient estimator. While the SNR of the standard Euclidean gradient estimator is known to vanish as the number of importance samples K increases, we prove that the SNR of the Wasserstein gradient scales favourably as $\Omega(\sqrt{K})$, ensuring optimisation efficiency even for large K. We further extend this geometric analysis to the Variational Rényi Importance-Weighted Autoencoder bound, establishing analogous stability guarantees. Experiments demonstrate that the proposed framework achieves superior approximation performance compared to other baselines.

IP20 — Statistical Estimation and Detection in Complex Modeling

June 16 (Tuesday) 13:30–15:10 **Room:** 209B Program

Haojie Ren, Shanghai Jiao Tong University, China

TBD

TBD

Shunxing Yan, Peking University, China

Semiparametric M-estimation with Overparameterized Neural Networks

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.

Xiaodong Yan, Xian Jiaotong University, China

AI Safety: Statistical Detection for Silent Data Corruption during Large-scale Model Training and Reasoning

uring large-scale model training, Silent Data Corruption (SDC) caused by hardware failures has emerged as a core challenge threatening model robustness and safety. SDC arises stochastically from factors such as chip transistor defects, voltage fluctuations, and temperature variations. Its concealment and latency make it difficult for traditional hardware detection methods, software redundancy techniques, and algorithm-level solutions to meet practical application standards in terms of coverage, computational overhead, and false alarm rate. To capture the system state transitions induced by SDC, this paper formulates the problem as Change-point Detection and constructs a novel test statistic based on the two-armed bandit framework. This approach achieves high performance (accuracy 99%, false alarm rate $< 10^{-8}$) with low performance loss ($<$ 2%). Simulation studies provide supporting evidence for the algorithm's performance, demonstrating its robustness with limited samples. Furthermore, a practical case study is included for illustrative purposes.

Xuening Zhu, Fudan University, China

TBD

TBD

IP51 — Advanced Machine Learning Methods for Challenges in Biomedical Data

June 16 (Tuesday) 13:30–15:10 **Room:** 203 Program

Changhu Wang, University of California, Los Angeles, United States

Nullstrap: A Simple, High-Power, and Fast Framework for FDR Control in Variable Selection for Diverse High-Dimensional Models

Balancing false discovery rate (FDR) control with high statistical power remains a central challenge in high-dimensional variable selection. While several FDR-controlling methods have been proposed, many degrade the original data—by adding knockoff variables or splitting the data—which often leads to substantial power loss and hampers detection of true signals. We introduce Nullstrap, a novel framework that controls FDR without altering the original data. Nullstrap generates synthetic null data by fitting a null model under the global null hypothesis that no variables are important. It then applies the

same estimation procedure in parallel to both the original and synthetic data. This parallel approach mirrors that of the classical likelihood ratio test, making Nullstrap its numerical analog. By adjusting the synthetic null coefficient estimates through a data-driven correction procedure, Nullstrap identifies important variables while controlling the FDR. We provide theoretical guarantees for asymptotic FDR control at any desired level and show that power converges to one in probability. Nullstrap is simple to implement and broadly applicable to high-dimensional linear models, generalized linear models, Cox models, and Gaussian graphical models. Simulations and real-data applications show that Nullstrap achieves robust FDR control and consistently outperforms leading methods in both power and efficiency.

Mladen Kolar, University of Southern California, United States

SMART: A Spectral Transfer Approach to Multi-Task Learning

Multi-task learning can substantially improve estimation by pooling information across related tasks, but its benefits often diminish in small-sample settings. A natural remedy is transfer learning, where information from a related source study is leveraged to aid a target problem. Most existing approaches rely on strong assumptions that the source and target models differ only in a bounded way, which can be overly restrictive in practice.

In this talk, I will introduce SMART, a spectral transfer method for multi-task linear regression that replaces bounded-difference assumptions with a more flexible notion of spectral similarity. We assume that the target coefficient matrix shares singular subspaces with the source, with alignment occurring sparsely. This structure arises naturally in many applications and enables effective transfer even when models differ substantially in magnitude.

SMART incorporates spectral information from a fitted source model through structural regularization, requiring no access to the source data. Despite the resulting nonconvex formulation, we develop an efficient ADMM-based algorithm with reliable performance guarantees. We establish non-asymptotic error bounds showing that SMART attains near minimax-optimal rates under mild conditions. Empirically, SMART delivers strong predictive gains, avoids negative transfer, and performs well on real multi-modal single-cell data.

Joint work with Boxin Zhao and Jinchi Lv.

Hongyuan Cao, Florida State University, United States

Statistical methods for fine mapping in admixed populations

Admixed populations provide unique opportunities for fine mapping because local-ancestry-informed genotypes capture distinct linkage disequilibrium patterns across ancestries, and causal variants may act in one ancestry or be shared across multiple ancestries. However, relatively few fine mapping methods are specifically designed for admixed populations, and existing approaches often assume ancestry-shared effects. To address this gap, we propose SuSiE-AP, a Bayesian fine mapping method tailored for admixed populations that allows genetic variants to have ancestry-specific or shared effects. Extensive simulations demonstrate that SuSiE-AP accurately localizes causal variants and distinguishes whether their effects are ancestry-specific or shared. This distinction enhances biological interpretation, clarifies disease mechanisms, and improves clinical translation and drug development. Overall, SuSiE-AP provides a scalable and interpretable framework for fine mapping in admixed populations.

Ran Dai, University of Nebraska Medical Center, United States

Controlling FDR in selecting group-level simultaneous signals from multiple data sources with application to the National COVID Collaborative Cohort data

One challenge in exploratory association studies using observational data is that the associations between the predictors and the outcome are potentially weak and rare, and the candidate predictors have complex correlation structures. False discovery rate (FDR) controlling procedures can provide important statistical guarantees for replicability in predictor identification in exploratory research. In the recently established National COVID Collaborative Cohort (N3C), electronic health record (EHR) data on the same set of candidate predictors are independently collected in multiple different sites, offering opportunities to identify true associations by combining information from different sources. This paper presents a general knockoff-based variable selection algorithm to identify associations from unions of group-level conditional independence tests (simultaneous signals) with exact FDR control guarantees under finite sample settings. This algorithm can work with general regression settings, allowing heterogeneity of both the predictors and the outcomes across multiple data sources. We demonstrate the performance of this method with extensive numerical studies and an application to the N3C data.

IP42 — Advances in Random Graph Theory

June 16 (Tuesday) 13:30–15:10 **Room:** 201 Program

Yuanfei Huang, Asia-Pacific Center for Theoretical Physics, South Korea

The SIR epidemic on a dynamic Erd{o}s-Rényi random graph

We investigate the SIR epidemic on a dynamic inhomogeneous Erd{o}s-Rényi random graph, in which edges appear and disappear independently of each other. We establish a functional law of large numbers for the susceptible, infected, and recovered ratio curves after a random time shift, and demonstrate that, under a variety of possible scaling limits of the model parameters, the epidemic curves are solutions to a system of ordinary differential equations. In most scaling regimes, these equations coincide with the classical SIR epidemic equations. There are two notable regimes where both the average degree of the network and the rate of the infectious transmission process remain constant. Furthermore, two novel sets of differential equations emerges. These systems contain additional quantities related to the infectious edges, but somewhat surprisingly, contain no quantities related to higher-order local network configurations. To the best of our knowledge, this study represents the first thorough and rigorous analysis of large population epidemic processes on dynamic random graphs, although our findings are contingent upon conditioning on a (possibly strict) subset of the event of an epidemic outbreak.

Xiao Fang, The Chinese University of Hong Kong, Hong Kong

Conditional central limit theorems for exponential random graphs

We study the Exponential Random Graph Models (ERGMs) conditioning on the number of edges. In subcritical region of model parameters, we prove a conditional Central Limit Theorem (CLT) with explicit mean and variance for the number of two stars. This generalizes the corresponding result in the literature for the Erdős–Rényi random graph. To prove our main result, we develop a new conditional CLT via exchangeable pairs based on the ideas of Dey and Terlov. Our key technical contributions in the application to ERGMs include establishing a linearity condition for an exchangeable pair involving two star counts, a local CLT for edge counts, as well as new higher-order concentration inequalities. Our approach also works for general subgraph counts, and we give a conjectured form of their conditional CLT. This talk is based on joint work with Song-Hao Liu, Zhonggen Su, Xiaolin Wang.

Rajat Hazra, Leiden University, The Netherlands

Voter Model on random graphs

In this talk, I will speak about the recent progresses in voter models on random graphs (both directed and undirected). We will study the consensus time in various setting and show how the heterogeneity of the degree sequence plays a crucial role in the consensus. We will also talk about some extensions to voter models with random rewiring dynamics. This is based on joint works with Luca Avena, Federico Capannoli, Frank den Hollander, Matteo Quattropani and Rangel Baldasso.

Gianmarco Bet, University of Florence, Italy

Localized geometry detection in scale-free random graphs

We consider the problem of detecting whether a power-law inhomogeneous random graph contains a geometric community, and we frame this as a hypothesis-testing problem. More precisely, we assume that we are given a sample from an unknown distribution on the space of graphs on n vertices. Under the null hypothesis, the sample originates from the inhomogeneous random graph with a heavy-tailed degree sequence. Under the alternative hypothesis, $k = o(n)$ vertices are given spatial locations and connect following the geometric inhomogeneous random graph connection rule. The remaining $n - k$ vertices follow the inhomogeneous random graph connection rule. We propose a simple and efficient test based on counting normalized triangles to differentiate between the two hypotheses. We prove that our test correctly detects the presence of the community with high probability as $n \to \infty$, and identifies large-degree vertices of the community with high probability.

CS08 — Conformal Prediction, Reinforcement Learning, and Modern Statistical Methods

June 16 (Tuesday) 13:30–15:10 **Room:** 202 Program

Chihoon Lee, Seoul National University

Predicting Current Outcomes From Historical Survey Data With Weighted Conformal Prediction

In large-scale complex surveys such as the National Health and Nutrition Examination Survey (NHANES), some outcomes are measured only in selected years, leaving incomplete records across survey waves. We develop a weighted conformal prediction framework that enables valid population-level prediction of unobserved outcomes using information from earlier surveys. The method accommodates covariate shift, where both continuous and categorical covariate distributions evolve over time while survey design affects representativeness. It integrates subgroup-specific density ratio and subgroup-proportion estimation to approximate likelihood ratios between the historical and target covariate distributions, and we establish coverage guarantees for the resulting prediction sets. Simulation studies and an application predicting low-density lipoprotein cholesterol (LDL-C) for the current U.S. population show that the proposed approach achieves coverage close to the nominal level and improved efficiency over existing methods, particularly when covariate distributions are complex or unknown.

Wenbo Jing, City University of Hong Kong

Knowledge Transfer in Batch Q Learning*

In data-driven decision-making across marketing, healthcare, and education, leveraging large datasets from existing ventures is crucial for navigating high-dimensional feature spaces and addressing data scarcity in new ventures. We investigate knowledge transfer in dynamic decision-making by focusing on batch stationary environments and formally defining task discrepancies through the framework of Markov decision processes (MDPs). We propose the Transfer Fitted Q-Iteration algorithm with general function approximation, which enables direct estimation of the optimal action-state function Q* using both target and source data. Under sieve approximation, we establish the relationship between statistical performance and the MDP task discrepancy, highlighting the influence of source and target sample sizes and task discrepancy on the effectiveness of knowledge transfer. Our theoretical and empirical results demonstrate that the final learning error of the function is significantly reduced compared to the single-task learning rate.

Soohyun Ahn, Ajou University

Conformal Monitoring of Complex Data with Finite-Sample Validity

This study proposes a conformal prediction-based monitoring framework for complex data, providing finite-sample Type I error control under exchangeability. By combining conformal prediction with kernel-based one-class classification, the proposed method flexibly accommodates structured and distributional data through appropriate distance or kernel choices. To illustrate the framework, we develop an SVDD-based chart with a Wasserstein kernel for histogram-valued data. Simulation studies and a real-data example demonstrate that the proposed method achieves accurate Type I error control and competitive detection power, with stable performance under moderate contamination.

Yongjae Kim, Seoul National University

An Association Measure for Mixed-Types Variables

Quantifying the association between a continuous variable and a categorical variable is a fundamental task in data analysis. Existing methods often rely on parametric assumptions or arbitrary integer encoding, which may lead to unstable results. We propose a distribution-free population measure of association, ξ', specifically designed for the mixed continuous-categorical setting. The proposed measure is normalized between 0 and 1; it equals 0 if and only if the variables are independent and 1 if and only if the categorical variable is a measurable function of the continuous one. We also introduce a corresponding sample estimator, ξ'_n, computable in $O(n \log n)$ time. These measures are invariant to permutations of

category labels and monotonic transformations of the continuous variable. We establish the strong consistency and asymptotic normality of the estimator ξ'_n, enabling a computationally efficient, permutation-free Wald test for independence, and an asymptotic confidence interval for the population measure ξ'. Extensive simulations and an application to The Cancer Genome Atlas (TCGA) data demonstrate that the proposed method yields superior power and stability compared to existing alternatives.

Tongyu Li, National University of Singapore

Statistical Inference on Gradient Flows

Gradient-based algorithms are central to modern statistical estimation, yet their statistical analysis is typically restricted to fixed-time behavior, such as convergence to a population truth or fluctuations at a prescribed iteration. In many applications, however, inference is required along the entire optimization path, particularly when the stopping time is data-dependent and diverging. In this paper, we develop a framework for time-uniform statistical inference on gradient flows arising from empirical risk minimization. We establish a functional central limit theorem that characterizes the deviation between the empirical and population gradient flows as a Gaussian process indexed by continuous time, uniformly over the nonnegative real line. Building on this result, we propose an algorithm-aware covariance estimation method that evolves jointly with the gradient flow, enabling computationally efficient estimation of the covariance structure without matrix inversion or resampling. We show that the proposed covariance estimator achieves consistency uniformly over time, which in turn facilitates the construction of confidence intervals for the target parameter with asymptotically valid coverage. Our results provide a unified perspective on optimization dynamics and statistical inference, and offer practical tools for uncertainty quantification in gradient-based methods.

CS12 — Biostatistics, Genomics, and Multiple Testing

June 16 (Tuesday) 13:30–15:10 **Room:** 214 Program

Hanning Chen, University of Melbourne

Moderated t-Covariates Improve Gene Ranking in Large-Scale Differential Expression Studies with Small Sample Sizes

Identifying differentially expressed (DE) genes between multiple experimental conditions is a fundamental problem in computational biology. However, two significant obstacles are often encountered: (1) identifying DE genes is particularly challenging in small-sample settings, and (2) in practice, researchers often have limited resources to verify a small number of genes.

Empirical Bayes approaches based on moderated genewise variances—most notably the limma framework (Smyth, 2004)—have proven effective in addressing these challenges. This project extends limma for large-scale studies with small sample sizes by introducing moderated t-covariates, motivated by multivariate t-distribution theory. These covariates are constructed from auxiliary contrasts that are orthogonal to the primary contrast of interest, thereby capturing additional information while remaining independent of the primary t-statistics under the null hypotheses.

Our approach provides a principled integration of two complementary lines of work: the moderated variance framework in limma for small-sample inference, and recent advances in covariate-assisted multiple testing that leverage auxiliary information to improve statistical power. The primary t-statistics and the moderated t-covariates are combined into a unified significance measure for gene prioritisation.

Simulation studies demonstrate that this construction maintains robust FDR control in small-sample settings, even when our t-distribution-based modelling assumptions are realistically violated. Furthermore, applications to RNA-seq and microarray datasets show that the proposed method effectively prioritises disease-related genes while increasing the number of discoveries, indicating more accurate detection.

Ninh Tran, University of Melbourne

A Covariate-Adaptive Test for Replicability Across Multiple Studies with False Discovery Rate Control

Replicability is a lynchpin for credible discoveries. The partial conjunction (PC) p-value, which combines individual base p-values from multiple similar studies, can gauge whether a feature of interest exhibits replicated signals across studies.

However, when a large set of features is examined, as in high-throughput experiments, testing for their replicated signals simultaneously can pose a very underpowered problem, due to both the multiplicity burden and inherent limitations of PC p-values. This power deficiency is markedly severe when replication is demanded for all studies under consideration, which is nonetheless the most natural and appealing benchmark for scientific generalizability a practitioner may request.

We propose ParFilter, a general framework that marries the ideas of filtering and covariate-adaptiveness to power up large-scale testing for replicated signals in the setting described above. It reduces the multiplicity burden by partitioning studies into smaller groups and borrowing the cross-group information to filter out unpromising features. Moreover, harnessing side information offered by auxiliary covariates whenever they are available, it can train informative hypothesis weights to encourage rejections of features more likely to exhibit replicated signals. We prove its finite-sample control on the false discovery rate, under both independence and arbitrary dependence among the base p-values across features. In simulations and two differential gene expression case studies, ParFilter demonstrates competitive performance relative to existing methodologies.

Armeen Taeb, University of Washington

Consensus Tree Estimation with False Discovery Rate Control via Partially Ordered Sets

Connected acyclic graphs (trees) are data objects that hierarchically organize categories. Collections of trees arise in a diverse variety of fields, including evolutionary biology, public health, machine learning, social sciences and anatomy. Summarizing a collection of trees by a single representative is challenging, in part due to the dimension of both the sample and parameter space. We frame consensus tree estimation as a structured feature-selection problem, where leaves and edges are the features. We introduce a partial order on leaf-labeled trees, use it to define true and false discoveries for a candidate summary tree, and develop an estimation algorithm that controls the false discovery rate at a nominal level for a broad class of nonparametric generative models. Furthermore, using the partial order structure, we assess the stability of each feature in a selected tree. Importantly, our method accommodates unequal leaf sets and non-binary trees, allowing the estimator to reflect uncertainty by collapsing poorly supported structure instead of forcing full resolution. We apply the method to study the archaeal origin of eukaryotic cells and to quantify uncertainty in deep branching orders. While consensus tree construction has historically been viewed as an estimation task, reframing it as feature selection over a partially ordered set allows us to obtain the first estimator with finite-sample and model-free guarantees. More generally, our approach provides a foundation for integrating tools from multiple testing into tree estimation.

Qin Shao, University of Toledo

Innovative Imputation Strategies for Time Series Data from Wearable Devices and Mobile Applications

One of the common problems of time series collected by wearable devices is incompleteness with high missing percentage. In addition, the missingness is more dominated by missing in group, where the values are missing at many successive time points. In this study, we aims to develop a Kalman filter based algorithm which enhances the accuracy of data analysis for datasets generated by the internet of medical things. The proposed algorithms for incomplete time series are compared on their out-of-sample predictive performances using simulated data sets of large and small sample sizes and with various missing rates. The consequence of model misspecification when the built-in correlations or autocorrelations of the time series are ignored is also studied. The application of the algorithm is demonstrated by analyzing a heart rate dataset.

CS14 — Latent Variable Models, Identifiability, and Complex Dependence Structures

June 16 (Tuesday) 13:30–15:10 **Room:** 215 Program

Chengyu Cui, University of Michigan

Beyond Vintage Rotation: Bias-Free Sparse Representation Learning with Oracle Inference

Learning low-dimensional latent representations is a central topic in statistics and machine learning, and rotation methods have long been used to obtain sparse and interpretable representations. Despite nearly a century of widespread use across

many fields, rigorous guarantees for valid inference for the learned representation remain lacking. In this paper, we identify a surprisingly prevalent phenomenon that suggests a reason for this gap: for a broad class of vintage rotations, the resulting estimators exhibit a non-estimable bias. Because this bias is independent of the data, it fundamentally precludes the development of valid inferential procedures, including the construction of confidence intervals and hypothesis testing. To address this challenge, we propose a novel bias-free rotation method within a general representation learning framework based on latent variables. We establish an oracle inference property for the learned sparse representations: the estimators achieve the same asymptotic variance as in the ideal setting where the latent variables are observed. To bridge the gap between theory and computation, we develop an efficient computational framework and prove that its output estimators retain the same oracle property. Our results provide a rigorous inference procedure for the rotated estimators, yielding statistically valid and interpretable representation learning.

Jiayi Huang, University of Virginia

How Does Missing Data Affect Latent Transition Analysis? A Monte Carlo Study

Latent Transition Analysis (LTA) is widely used to model longitudinal changes in unobserved categorical states, yet its performance can be highly sensitive to missing data, a pervasive issue in applied research. Despite its importance, there is limited guidance on how different missing data mechanisms and analytic approaches affect the recovery of latent classes and transition dynamics. This study addresses this gap through a Monte Carlo simulation of a two-class LTA model with binary indicators, systematically varying sample size, number of time points, number of indicators, missing data mechanism (MCAR vs. MAR), and missing data rate. Models are estimated using Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI).

Results show that parameter recovery is consistently more accurate under MCAR than MAR missingness. FIML demonstrates robust performance across a wide range of conditions, while MI is more sensitive to model complexity and missingness patterns. Importantly, increasing sample size and number of indicators substantially improves latent class recovery and reduces bias in both emission and transition parameters.

These findings provide clear, practical guidance on how study design and missing data mechanisms jointly influence LTA performance, and offer actionable recommendations for applied researchers working with longitudinal data.

Chengzhu Huang, Columbia University

A General Recipe for Generalized Latent Factor Models: Missingness, Implicit Regularization, and Inference

Generalized latent factor models provide a flexible framework for analyzing high-dimensional discrete data, yet inference procedures under missingness remain largely underdeveloped. We study nonlinear latent factor models with exponential-family links and partially observed entries, and propose a unified computational and inferential pipeline that is both statistically optimal and algorithmically tractable. Our method proceeds in three stages: soft singular value thresholding for initialization, a one-step refinement that attains row-wise consistency, and a vanilla gradient descent scheme that exploits implicit regularization to navigate the nonconvex likelihood landscape without explicit penalties. We show that, under suitable initialization and learning rates, gradient descent converges rapidly while preserving key identifiability structures. Leveraging a linear approximation of the iterates, we construct valid individual and simultaneous confidence bands for latent factors and intercepts via Gaussian multiplier bootstrap. To our knowledge, this is the first framework that provides explicit guidance for initializing gradient descent and demonstrates that such properly initialized procedures can directly enable statistical inference in nonlinear low-rank models with missing data. Extensive simulations and real data analysis confirm accurate estimation, reliable coverage, and robustness to missingness, yielding a practical toolkit for uncertainty quantification in high-dimensional latent factor analysis.

Mengqi Lin, University of Michigan

Characterizing Identifiability in Boolean Graphical Models

Boolean graphical models, including prominent subfamilies such as Boolean matrix decompositions and cognitive diagnosis models, find broad applications ranging from social sciences to engineering. Despite their flexibility, a key challenge lies in establishing the identifiability of their graphical structures, which specify how latent variables influence observed variables. Existing identifiability conditions typically rely on the strong assumption of pure nodes, which may be unrealistic in many applications. We develop a novel approach leveraging the Hasse diagram to represent the distribution of

observed variables and transform identifiability into a graph isomorphism challenge. Based on this, we establish *sufficient and necessary* graphical identifiability conditions that do not require pure nodes. We further derive equivalent algebraic conditions and develop an efficient Boolean satisfiability (SAT)-based verification algorithm. Our results substantially broaden the class of identifiable and interpretable Boolean graphical models by removing the pure-node requirement, yielding new theoretical insights, while also providing practitioners with a concrete and easily implementable tool to assess model identifiability.

Jimin Kim, Seoul National University

Conditional Copula Networks for Synthetic Tabular Data with Complex Dependencies

This study proposes a novel generative model, termed the Conditional Copula Network (CCN), for synthetic tabular data generation that explicitly models inter-variable dependencies in mixed-type data. While recent deep learning-based approaches have demonstrated strong performance in capturing marginal distributions, they often fail to preserve the underlying dependency structure, leading to distortions in correlations. To address this limitation, we adopt a copula-based statistical framework that explicitly separates marginal distributions from dependency structures, and extend it by integrating a neural architecture.

The proposed CCN maps categorical configuration into a continuous embedding space and learns the conditional covariance structure of numerical variables given these configuration. To balance global consistency and local flexibility, we introduce a shared-differential decomposition of the covariance matrix, which captures both global dependency patterns and configuration-specific variations. Furthermore, a Cholesky-based parameterization is employed to ensure positive definiteness and numerical stability of the covariance estimates.

We evaluate the proposed model on several real-world tabular datasets using the TSTR(Train on Synthetic, Test on Real) framework. Experimental results demonstrates that CCN achieves superior performance, particularly in regression task, by effectively reconstructing conditional dependencies among numerical variables. In addition, the model exhibits strong robustness in reproducing diverse distributions and maintains stable performance across mixed-type data settings.

Overall, CCN provides a principled and interpretable approach to synthetic tabular data generation by combining the statistical foundation of copulas with the expressive power of deep learning

Author Index

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z