HKU 2025 Summer Workshop on Statistics and Data Analytics

Sessions:

Venue: CPD-3.28, Central Podium Levels – Three, The Jockey Club Tower, Centennial Campus, HKU.

Time	People	Titles/Activities
08:30–09:00	Registration	
09:00–09:10	Opening: Haipeng Shen & Zhixi Wan	
Session 1	**Chair: Haipeng Shen**	
09:10–09:50	Tony Cai	Transcending Data Boundaries: Transfer Knowledge in Statistical Learning
09:50–10:00	Photo	
10:00–10:30	Coffee break	
Session 2	**Chair: Dan Yang & Xinghao Qiao**	
10:30–11:10	Rong Chen	Dynamic Tensor Factor Model with Main and Interaction Effects
11:10–11:50	Junhui Wang	Learning Nonparametric Graphical Model on Heterogeneous Network-linked Data
11:50–14:00	Lunch Break	
Session 3	**Chair: Jing Ouyang & Weichen Wang**	
14:00–14:40	Yingying Fan	Asymptotic FDR Control with Model-X Knockoffs: Is Moments Matching Sufficient?
14:40–15:20	Yongmiao Hong	Reinforced Tail Quantile Regression
15:20–15:50	Coffee break	
Session 4	**Chair: Zhanrui Cai & Xin Tong**	
15:50 - 16:30	Xi Chen	LLM Alignment Techniques: Stochastic Optimizations in LLM Post-training and Reasoning.
16:30 - 17:00	Gareth James	The Role of Statistics Within Business in the Era of AI
17:00 - 17:10	Coffee Break	
Panel Discussion: Development of Statistics within Business School		
17:10 - 17:55	Panel Discussion	
17:55 - 18:00	Closing: Xinghao Qiao & Weichen Wang	
18:00 - 20:30	Banquet (invitation only)	

For detailed direction to the venues, please visit https://hkubs-stat.github.io/HKUBA-2025-Summer-Workshop/

HKU BUSINESS SCHOOL
港大經管學院

Abstract

1. **Transcending Data Boundaries: Transfer Knowledge in Statistical Learning**

   Speaker: Tony Cai, University of Pennsylvania

   Abstract: Human learners have a natural ability to transfer knowledge from one setting to another, using experience gained in one context to inform learning in a different but related one. This capacity for transfer is fundamental to efficient and adaptive learning. In contrast, most statistical learning procedures are designed to address a single task or learn a single distribution, based solely on data from that specific setting.

   In this talk, I will discuss recent optimality results in statistical transfer learning in various settings, with a primary focus on functional mean estimation and covariance matrix estimation. These results demonstrate the significant benefits of incorporating data from source distributions to enhance learning performance under the target distribution.

2. **Dynamic Tensor Factor Model with Main and Interaction Effects**

   Speaker: Rong Chen, Rutgers University

   Abstract: High dimensional tensor time series has been encountered increasingly often in applications. Factor model in a form similar to tensor Tucker decomposition has been shown to be a useful model for tensor time series. In this paper we propose a more detailed decomposition so the factors can be interpreted as global effects, main effects of individual modes (columns, rows, etc), and interaction effects among the modes. This decomposition enhances interpretability, effective dimension reduction and estimation efficiency. Theoretical investigation establishes the properties of the estimation procedure. Empirical examples are used to illustrate the applicability of the methodology, highlighting its relevance to contemporary data science challenges in high-dimensional settings.

3. **LLM Alignment Techniques: Stochastic Optimizations in LLM Post-training and Reasoning.**

   Speaker: Xi Chen, New York University

   Abstract: This talk explores approaches to improving large language model (LLM) post-training and reasoning through stochastic optimization techniques. The first part introduces ComPO, a preference alignment method using comparison oracles in stochastic optimization. The work addresses likelihood displacement issues in traditional direct preference optimization. The second part proposes the spectral policy optimization, a framework that overcomes GRPO's limitations with all-negative-sample groups by introducing response diversity with AI feedback. Both approaches demonstrate significant improvements across various model sizes and benchmarks, representing important advances in LLM post-training via stochastic optimization.

4. Asymptotic FDR Control with Model-X Knockoffs: Is Moments Matching Sufficient?

Speaker: Yingying Fan, University of Southern California

Abstract: We propose a unified theoretical framework for studying the robustness of the model-X knockoffs framework by investigating the asymptotic false discovery rate (FDR) control of the practically implemented approximate knockoffs procedure. This procedure deviates from the model-X knockoffs framework by substituting the true covariate distribution with a user-specified distribution that can be learned using in-sample observations. By replacing the distributional exchangeability condition of the model-X knockoff variables with three conditions on the approximate knockoff statistics, we establish that the approximate knockoffs procedure achieves the asymptotic FDR control. Using our unified framework, we further prove that an arguably most popularly used knockoff variable generation method--the Gaussian knockoffs generator based on the first two moments matching--achieves the asymptotic FDR control when the two-moment-based knockoff statistics are employed in the knockoffs inference procedure. For the first time in the literature, our theoretical results justify formally the effectiveness and robustness of the Gaussian knockoffs generator. Simulation and real data examples are conducted to validate the theoretical findings.

5. Reinforced Tail Quantile Regression

Speaker: Yongmiao Hong, University of Chinese Academy of Sciences

Abstract: Quantile regression in the tails suffers from inconsistency and a non-normal asymptotic distribution due to data sparsity. To address this situation, we propose a reinforced tail quantile regression estimator that leverages the power-law behavior for heavy-tailed data. Our estimator is both consistent and asymptotically normal under some regularity conditions. Furthermore, we establish the asymptotic validity of bootstrap inference using random weights. Simulation results demonstrate the superior performance of our estimator and the near-exact coverage of the bootstrap confidence intervals. In particular, our method yields narrower confidence intervals compared to existing approaches. We apply the proposed method to examine the marginal effect of education on the upper extreme percentiles of income, using a unique dataset from the Chinese Twins Survey conducted by the National Bureau of Statistics. The sample includes 2,412 individuals, with an average monthly income of CNY 912 in 2002. The income distribution exhibits heavy-tailed behavior, with a tail index estimated at 2, implying that the moments of order higher than two may not exist. Our method uncovers a significantly positive effect of education in the tail, in contrast to existing approaches, which yield insignificant or even negative effects, often accompanied by wide confidence intervals. Notably, the width of the confidence intervals from our method remains stable in the tail region and is comparable to that of the standard quantile regression at fixed quantile levels.

6. The Role of Statistics Within Business in the Era of AI

Speaker: Gareth James, Emory University

Abstract: The 2010's saw a dramatic increase in the availability of business data and a corresponding rise in the importance of data science and machine learning. As a result, we have seen significant growth in the numbers of statisticians and computer scientists within business schools. Arguably, the first half of the 2020's has seen an even more dramatic revolution, with AI taking center stage. While we are all trying to assess the long-term impacts of AI, it is clear that these Large Language Model (LLM) technologies pose both opportunities and challenges for statisticians. In this seminar I will present some of my perspectives on this topic as a statistician with over 25 years' experience teaching in, and leading, business schools.

7. Learning Nonparametric Graphical Model on Heterogeneous Network-linked Data

Speaker: Junhui Wang, The Chinese University of Hong Kong

Abstract: Graphical models have been popularly used for capturing conditional independence structure in multivariate data, which are often built upon independent and identically distributed observations, limiting their applicability to complex datasets such as network-linked data. In this talk, we introduce a nonparametric graphical model that addresses these limitations by accommodating heterogeneous graph structures without imposing any specific distributional assumptions. The introduced estimation method effectively integrates network embedding with nonparametric graphical model estimation. It further transforms the graph learning task into solving a finitedimensional linear equation system by leveraging the properties of vectorvalued reproducing kernel Hilbert space. We will also discuss theoretical properties of the proposed method in terms of the estimation consistency and exact recovery of the heterogeneous graph structures. Its effectiveness is also demonstrated through a variety of simulated examples and a real application to the statistician coauthorship dataset.