

The Second

HONG KONG-MACAO-SHANGHAI

Symposium on Statistics and Data Science

PROGRAM BOOK

Date:
14 November 2026 (Saturday)

Time:
08:55 – 17:20

Venue:
LT4, 2/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong

Organized by
Department of Statistics and Data Science,
The Chinese University of Hong Kong

Program at a Glance

Saturday, 14 November 2026 · LT4, YIA, CUHK · 20 min per talk incl. Q&A · Tentative

08:30 – 08:55	**Registration**	
08:55 – 09:00	**Opening Ceremony + Group Photo** Opening speeches: ZHANG Riquan (SUIBE) and ZHANG Wenyang (University of Macau) Moderator: Zhixiang LIN	
09:00 – 10:20	**Session 1** Session Chair: Zhixiang LIN	
09:00	**Zhi LIU** University of Macau	*Nonparametric Test for Diurnal Variation in Intraday Correlation Processes*
09:20	**Hongmei LIN** Shanghai University of International Business and Economics	*A Transfer Learning Framework for Efficient Portfolio Selection*
09:40	**Long YU** Shanghai University of Finance and Economics	*High-Dimensional Factor and Clustering Analysis via Projected and Truncated Sample Covariance Matrix*
10:00	**Mengying YOU** Shanghai University of International Business and Economics	*Neural Spatial Dimension Reduction and State-Space Forecasting*
10:20 – 10:40	**Coffee Break** Tea reception area near YIA LT4	
10:40 – 12:00	**Session 2** Session Chair: Ben DAI	
10:40	**Qian XIAO** Shanghai Jiao Tong University	*Optimal Design of Experiments with Quantitative-Sequence Factors*
11:00	**Le ZHOU** Hong Kong Baptist University	*Rank-Based Sparse Regression in Principal Components Space under Measurement Error*
11:20	**Biao CAI** City University of Hong Kong	*A Preferential Latent Space Model for Text Networks*
11:40	**Jiyuan TU** Shanghai University of Finance and Economics	*Sub-Gradient Methods for Non-Smooth Empirical Loss Minimization*
12:00 – 14:00	**Lunch** By invitation	
14:00 – 15:20	**Session 3** Session Chair: Yingxin LIN	
14:00	**Haojie REN** Shanghai Jiao Tong University	*Generalized Boundary Error Control in Multiple Testing*
14:20	**Yingxin LIN** The Chinese University of Hong Kong	*Spatially Structured Data Splitting for False Discovery Control in Spatial-Domain Marker Discovery*
14:40	**Yajie BAO** Fudan University	*Conformal Prediction on Constrained Spaces: A General Projection Framework*
15:00	**Rui QIU** East China Normal University	*Metric Conformal Prediction via the Conditional Expected Local Radius*
15:20 – 15:40	**Coffee Break** Tea reception area near YIA LT4	
15:40 – 17:20	**Session 4** Session Chair: Gen LI	
15:40	**Guanhua FANG** Fudan University	*Large Language Model Reasoning under Token Budgets: A Competing Risks Analysis*
16:00	**Rong TANG** The Hong Kong University of Science and Technology	*Neural Estimation of Summary Statistics for Simulation-based Inference with Non-asymptotic Guarantees*
16:20	**Gen LI** The Chinese University of Hong Kong	*Sampling Acceleration of Diffusion Language Models*
16:40	**Jin-Hong DU** The University of Hong Kong	*When LLMs Judge LLMs: Disagreement, Position Bias, and Human Calibration*
17:00	**Yuheng MA** East China Normal University	*Black-Box Test-Time Adaptation: Optimality, Personalization and Collaboration*
18:00	**Banquet** By invitation	

Session 1

09:00 – 10:20 | Session Chair: Zhixiang LIN

09:00 – 09:20

Zhi LIU

Department of Mathematics, University of Macau

Nonparametric Test for Diurnal Variation in Intraday Correlation Processes

The association between log-price increments of exchange-traded equities, as measured by their spot correlation estimated from high-frequency data, exhibits a pronounced upward-sloping and almost piecewise linear relationship at the intraday horizon. There is notably lower---on average less positive---correlation in the morning than in the afternoon. We develop a nonparametric testing procedure to detect such variation in a correlation process. The test statistic has a known distribution under the null hypothesis, whereas it diverges under the alternative. We run a Monte Carlo simulation to discover the finite sample properties of the test statistic, which are close to the large sample predictions, even for small sample sizes and realistic levels of diurnal variation. In an application, we implement the test on a high-frequency dataset covering the stock market over an extended period. The test leads to rejection of the null most of the time. This suggests diurnal variation in the correlation process is a nontrivial effect in practice. We show how conditioning information about macroeconomic news and corporate earnings announcements affect the intraday correlation curve.

09:20 – 09:40

Hongmei LIN

School of Statistics and Data Science, Shanghai University of International Business and Economics

A Transfer Learning Framework for Efficient Portfolio Selection

Portfolio selection plays a critical role in financial decision-making by helping investors optimize returns, meet financial objectives, and manage risk. However, practical applications often face challenges such as limited data and high noise levels, particularly in emerging markets. This paper investigates the potential of transfer learning to address these issues in portfolio selection. We propose a general framework that incorporates information from auxiliary sources through a proposed transferable source detection algorithm. We further provide a rigorous theoretical analysis, establishing ℓ_1-norm estimation error bounds and consistency guarantees for the proposed methods. Empirical results on both simulated and real-world datasets demonstrate that our transfer learning approach substantially improves portfolio performance, yielding higher Sharpe ratios and more efficient asset allocation, especially when more source data are available or a logarithmic utility function is employed.

09:40 – 10:00

Long YU

School of Statistics and Data Science, Shanghai University of Finance and Economics

High-Dimensional Factor and Clustering Analysis via Projected and Truncated Sample Covariance Matrix

This paper addresses spike identification in high-dimensional mixture factor models, where the spiked eigenvalues of the covariance matrix may arise from clustered mean vectors and/or latent common factors. The magnitudes of the eigenvalues from these two sources can be comparable, posing a significant challenge to model identification and estimation. We introduce a novel projected and truncated sample covariance matrix which, with suitable thresholding, consistently recovers the factor loading space regardless of the relative strength of the cluster-driven eigenvalues. Importantly, this consistency holds even when the subspaces induced by the factor structure and the clustered means are partially confounded. Given this estimated factor space, we propose a projected spectral clustering procedure that removes the factor component and reveals the underlying cluster structure. Furthermore, we develop an iterative refinement algorithm that builds upon these steps to further enhance estimation accuracy. All proposed methods are supported by rigorous theoretical guarantees and are validated through extensive simulations and real data analyses.

10:00 – 10:20

Mengying YOU

School of Statistics and Data Science, Shanghai University of International Business and Economics

Neural Spatial Dimension Reduction and State-Space Forecasting

Forecasting high-dimensional spatial fields requires effective dimension reduction and a model for temporal evolution. We propose a two-stage framework combining neural spatial representation learning with state-space forecasting. In the first stage, a neural network learns a low-rank representation through a regularized reconstruction objective. The learned spatial loadings are orthogonalized to define a fixed coordinate system, and the observed fields are projected onto this system to obtain temporal scores. In the second stage, state-space models describe score dynamics and generate forecasts, which are mapped back to the spatial observation scale. Representation learning and temporal modeling are estimated separately, allowing spatial structure and temporal dynamics to be modeled within distinct stages. We illustrate the framework using county-level COVID-19 observations in the United States, examining spatial reconstruction, short-horizon forecasting, and predictive uncertainty. The analysis highlights the distinction between reconstruction quality and forecasting performance, as well as between point accuracy and predictive density assessment. The framework connects flexible neural dimension reduction with statistical temporal modeling for high-dimensional spatiotemporal data.

Session 2

10:40 – 12:00 | Session Chair: Ben DAI

10:40 – 11:00

Qian XIAO

School of Mathematical Sciences, Shanghai Jiao Tong University

Optimal Design of Experiments with Quantitative-Sequence Factors

A new type of experiments with joint considerations of quantitative and sequence factors are recently drawing much attention in medical science, bio-engineering and many other disciplines. The input spaces of such experiments are semi-discrete and often very large. Thus, efficient and economic experimental designs are required. Based on the transformations and aggregations of good lattice point sets, we construct a new class of optimal quantitative-sequence (QS) designs which are marginally coupled, pair-balanced, space-filling and asymptotically orthogonal. The proposed QS designs have certain flexibility in run and factor sizes, and are especially appealing for high-dimensional cases.

11:00 – 11:20

Le ZHOU

Department of Mathematics, Hong Kong Baptist University

Rank-Based Sparse Regression in Principal Components Space under Measurement Error

We study high-dimensional regression in principal components space when the predictors are observed with additive measurement error and the response errors may be heavy-tailed. The starting point is the ℓ_1-penalized principal-components estimator of Song and Zou (2026), which enjoys a blessing-of-dimensionality phenomenon under predictor contamination but is sensitive to heavy-tailed data or outliers. To address this issue, we propose a two-stage rank-based estimator that combines Wilcoxon rank loss with adaptive reweighting through a folded-concave penalty. The proposed method is robust to heavy-tailed response errors while preserving the principal-components structure induced by contaminated predictors. We further establish a nonasymptotic prediction bound for the second-stage fitted mean, and show that, most importantly, the blessing-of-dimensionality phenomenon survives in this setting. Simulation studies demonstrate that our proposed method remains competitive under Gaussian noise and is substantially more stable under heavy-tailed errors, especially when predictor contamination is present.

11:20 – 11:40

Biao CAI

Department of Decision Analytics and Operations, City University of Hong Kong

A Preferential Latent Space Model for Text Networks

Network data enriched with textual information, referred to as text networks, arise in a wide range of applications, including email communications, scientific collaborations, and legal contracts. In such settings, both the structure of interactions (i.e., who connects with whom) and their content (i.e., what is communicated) are useful for understanding network relations. Traditional network analyses often focus only on the structure of the network and discard the rich textual information, resulting in an incomplete or inaccurate view of interactions. In this paper, we introduce a new modeling approach that incorporates texts into the analysis of networks using topic-aware text embedding, representing the text network as a generalized multilayer network where each layer corresponds to a topic extracted from the data. We develop a new and flexible latent space network model that captures how node-topic preferences directly modulate edge formation, and establish identifiability conditions for the proposed model. We tackle model estimation with a projected gradient descent algorithm, and further discuss its theoretical properties. The efficacy of our proposed method is demonstrated through simulations and an analysis of an email network.

11:40 – 12:00

Jiyuan TU

School of Statistics and Data Science, Shanghai University of Finance and Economics

Sub-Gradient Methods for Non-Smooth Empirical Loss Minimization

In this paper, we investigate the optimization of non-smooth empirical loss functions from a statistical perspective. Viewing the empirical loss as a non-smooth approximation to an underlying smooth population loss, we introduce three conditions—approximate smoothness, approximate strong convexity, and local smoothness—that characterize its geometry at different scales. Under these conditions, we show that classical first-order methods, including sub-gradient descent and Nesterov's accelerated sub-gradient descent, can achieve convergence behavior comparable to their smooth-loss counterparts. This leads to a three-stage convergence framework: the algorithms initially converge at a polynomial rate, as in smooth convex optimization; after entering a locally strongly convex region, they converge geometrically; and, in the final local regime, the iterates admit a Bahadur-type representation with optimization error negligible relative to statistical error. We further study two representative smoothing techniques, infimum convolution and integral convolution, and show that appropriate small-scale smoothing can upgrade approximate smoothness and strong convexity to their exact counterparts, substantially improving algorithmic stability near the target. Finally, we extend the framework to high-dimensional sparse problems by introducing restricted versions of the proposed nonsmoothness conditions. For proximal sub-gradient descent and iterative hard thresholding, we establish analogous stage-wise convergence behavior, showing that these methods can retain smooth-like optimization properties while delivering statistically valid post-stage estimators. Simulation studies confirm the effectiveness of the proposed methods.

Session 3

14:00 – 15:20 | Session Chair: Yingxin LIN

14:00 – 14:20

Haojie REN

School of Mathematical Sciences, Shanghai Jiao Tong University

Generalized Boundary Error Control in Multiple Testing

False discovery rate control safeguards the average quality of a discovery set but need not ensure the reliability of its least significant discoveries. These boundary discoveries are often the most uncertain findings, yet their reliability can be crucial to subsequent scientific investigation. This talk presents generalized boundary error control in offline and online multiple testing. The central criterion is the k-boundary false discovery rate, defined as the probability that the k least significant discoveries are all false. We develop a closure-based framework that focuses on certifying the rejection boundary rather than protecting the entire rejection set against any k false discoveries. This approach yields procedures that identify valid boundaries in offline analyses and preserve boundary certification as hypotheses arrive in online studies. Using appropriately constructed local intersection tests, the proposed procedures provide finite-sample error control under arbitrary dependence, with computationally efficient implementations available under suitable structural conditions. Simulations and real-data analyses illustrate how this perspective supports reliable boundary discoveries while retaining useful discovery power.

14:20 – 14:40

Yingxin LIN

Department of Statistics and Data Science, The Chinese University of Hong Kong

Spatially Structured Data Splitting for False Discovery Control in Spatial-Domain Marker Discovery

Spatial omics technologies profile molecular measurements while preserving tissue architecture, but valid marker discovery after unsupervised spatial clustering remains statistically challenging. Standard workflows first infer spatial domains and then use the same observations for differential expression across domains. This double use of the same data can inflate false discoveries, especially when spatial boundaries are weak or neighboring observations are correlated. We introduce SpaMDS, a spatially aware multiple data splitting framework for false discovery rate controlled marker discovery after spatial clustering. SpaMDS divides data into two subsets using spatially structured splits, applies the same clustering and differential expression pipeline to each subset, and combines signed feature statistics through mirror statistics across repeated splits. In simulations based on spatial transcriptomics data, SpaMDS maintained nominal FDR control while preserving marker recovery across two-domain and three-domain settings, different spatial geometries, domain-size ratios, marker-set sizes, and signal strengths. The gains were most apparent in low-signal settings where existing procedures were either overly permissive or conservative. In DLPFC Visium data with manual anatomical annotations, SpaMDS achieved a better balance between marker recovery and false positive control. In Visium HD human colorectal cancer data, SpaMDS reduced marker selection for cluster comparisons with limited biological specificity while retaining biologically interpretable tumor-domain markers. These results support SpaMDS as a practical strategy for post-clustering marker discovery in spatial omics.

14:40 – 15:00

Yajie BAO

Department of Statistics and Data Science, Fudan University

Conformal Prediction on Constrained Spaces: A General Projection Framework

In many applications, prediction sets must satisfy known structural constraints while providing reliable coverage of future responses. This work develops a general framework for constructing such sets through projection and conformal calibration. Starting from an initial family of prediction sets, we project entire sets onto the constraint region and recalibrate the resulting family to reduce conservatism. For any pretrained nonconformity score, we express the corresponding projected score as the optimal value of a constrained optimization problem. For commonly used multivariate prediction-set shapes, this problem admits a convex formulation under general convex constraints. Our theoretical analysis characterizes when the proposed method yields smaller prediction sets than those obtained by directly intersecting standard conformal sets with the constraint region. We further optimize the projection rule to minimize the volume of the projected prediction sets while retaining finite-sample coverage guarantees.

15:00 – 15:20

Rui QIU

School of Statistics, East China Normal University

Metric Conformal Prediction via the Conditional Expected Local Radius

This paper studies prediction for non-Euclidean responses in general metric spaces, where a vector space structure and a natural density need not be available. We develop a conformal method based on the conditional expected local radius (CELR), the conditional analogue of the L^1 distance-to-a-measure. At each candidate response, CELR averages the radii needed to accumulate conditional probability mass over levels up to a specified threshold. Transforming CELR by its conditional distribution gives a nonconformity score that is comparable across covariate values and requires neither coordinates nor densities. A metric distributional random forest estimates the conditional law with multivariate covariates, using the same forest weights to estimate distance profiles, local radii, and their conditional distributions. We prove uniform consistency of these estimators and asymptotic conditional validity of the resulting prediction sets. Split conformal calibration provides finite-sample marginal coverage under exchangeability. Numerical studies demonstrate effective uncertainty quantification with multivariate covariates across different metric spaces.

Session 4

15:40 – 17:20 | Session Chair: Gen LI

15:40 – 16:00

Guanhua FANG

Department of Statistics and Data Science, Fudan University

Large Language Model Reasoning under Token Budgets: A Competing Risks Analysis

Assessing language-model performance under a token budget requires understanding both when a response will finish and whether it will be correct. We develop CRATER — Competing Risks Analysis of Token Expenditure in Reasoning — a framework that treats correct and incorrect completion as distinct event types and responses unfinished at the generation limit as administratively right-censored. The fitted distribution provides probabilities of correct completion, incorrect completion, and remaining unfinished across token budgets, together with expected token consumption. We evaluate the proposed method using responses to MATH questions and illustrate its use in two downstream applications. For routing before generation, CRATER guides the joint choice of a model configuration and token budget. Under hypothetical token prices, it reduces mean cost by 53.39% relative to a strong reference, with observed correct-completion rates of 96.42% and 97.30%, respectively. For repeated sampling, we combine CRATER's single-attempt success probabilities with an existing beta-binomial calibration approach to predict pass@k. Across both applications, CRATER provides predictions at any specified token budget up to the generation limit from a common model of completion time and correctness, without requiring separate parameter estimation for each budget.

16:00 – 16:20

Rong TANG

Department of Mathematics, The Hong Kong University of Science and Technology

Neural Estimation of Summary Statistics for Simulation-based Inference with Non-asymptotic Guarantees

Simulation-based inference (SBI) enables Bayesian inference when a simulator is available but the likelihood is intractable. Modern SBI methods train neural networks on simulated data to obtain an amortized approximation of the posterior, making inference for new observations fast. A central challenge in SBI is to extract informative, low-dimensional summary statistics; without them, performance can degrade rapidly as the observation dimension grows. We therefore highlight summary learning as a critical component of SBI and propose a general framework that jointly learns a neural-network–parameterized summary and an amortized posterior estimator. We further introduce a local refinement strategy that adapts the amortized estimator to a specific observation, thereby reducing the amortization gap. Both amortized training and local refinement accommodate implicit posterior estimators without tractable densities. On the theoretical side, we analyze the finite-sample behavior of the joint learning procedure through the lens of nonparametric conditional distribution estimation, and show on representative examples how low-dimensional summaries can mitigate the classical curse of dimensionality. Empirically, our results support this motivation and demonstrate favorable performance relative to existing methods.

16:20 – 16:40

Gen LI

Department of Statistics and Data Science, The Chinese University of Hong Kong

Sampling Acceleration of Diffusion Language Models

Diffusion language models (DLMs) have emerged as a compelling alternative to autoregressive (AR) models by enabling parallel, non-sequential token generation. Despite their remarkable empirical success, the theoretical foundations of these models remain relatively underexplored, particularly in terms of convergence guarantees. In this talk, we will present recent theoretical convergence guarantees for both discrete and continuous DLMs. These results provide a theoretical foundation for efficient diffusion-based language generation and offer principled insights into decoding strategy design.

16:40 – 17:00

Jin-Hong DU

Department of Statistics and Actuarial Science, The University of Hong Kong

When LLMs Judge LLMs: Disagreement, Position Bias, and Human Calibration

Large language models (LLMs) are increasingly used as judges to compare responses and rank other LLMs. For statisticians, they represent a new kind of rater: abundant and inexpensive, but of uncertain reliability. Real benchmarks such as Chatbot Arena reveal three challenges: LLM judges disagree systematically, their judgments depend on response display order, and their preferences can diverge from those of humans. This talk addresses these challenges through two complementary pairwise comparison frameworks. First, we decompose judges' latent preference scores into a shared consensus direction and low-rank, judge-specific deviations, separating sensitivity to consensus from systematic disagreement, with identification and inference guarantees. Second, building on this structure, we estimate and remove judge-specific position effects and calibrate the resulting LLM preference space toward latent human preferences using limited human comparisons. Adaptive weighting balances information from LLM and human judgments, allowing the calibration to adjust when LLM preferences are imperfect. On Chatbot Arena, as few as 50 human comparisons yield rankings closely aligned with held-out human preferences, while adaptive calibration shifts reliance toward human evidence when the LLM representation is less informative. Together, these results illustrate how abundant but imperfect LLM judgments can support statistically principled, human-aligned evaluation.

17:00 – 17:20

Yuheng MA

School of Statistics, East China Normal University

Black-Box Test-Time Adaptation: Optimality, Personalization and Collaboration

Test-time adaptation offers a flexible way to personalize large language models (LLMs) without modifying their parameters, but becomes challenging when the model is accessible only as a black box and no reliable reward model or verifier is available. We study this problem under the Best-of-N framework: at each interaction, a frozen language model generates a set of candidate responses, the system selects one to present to the user, and feedback is observed only for the selected response. We develop AdaptiveNN, an adaptive nearest-neighbor UCB procedure that learns user preferences directly from such bandit feedback. Under mild regularity conditions, we establish matching regret upper and minimax lower bounds up to logarithmic factors. The minimax rate exhibits a phase transition in the number of generated candidates, quantifying the statistical trade-off between generating more alternatives and learning to select among them. We then extend the framework to collaborative personalization across heterogeneous users. We introduce a winner-based notion of preference similarity that compares users through their preferred choices over random candidate sets and is invariant to user-specific monotone reward calibrations. This structure leads to a collaborative procedure that identifies compatible users and pools their feedback while accounting for heterogeneous reward calibrations. We establish guarantees for cluster recovery, collaborative regret, and imperfect clustering, and characterize how the benefits of collaboration are distributed across users. Experiments on synthetic problems and question-answering benchmarks support the theoretical findings and illustrate the potential of bandit feedback for personalized black-box adaptation.

General Information

Date	14 November 2026 (Saturday)
Venue	LT4, 2/F, Yasumoto International Academic Park (YIA), The Chinese University of Hong Kong
Registration	08:30 – 08:55
Coffee breaks	10:20 – 10:40 and 15:20 – 15:40, tea reception area near YIA LT4
Lunch	12:00 – 14:00 (by invitation)
Banquet	18:00 (by invitation)
Organizer	Department of Statistics and Data Science, The Chinese University of Hong Kong