Invited Sessions III - 12 December 2024, 09:00 – 10:30

Session 03	**Advances in Statistical and Machine Learning Methods for Complex Dependent Data** *Organizer/Chair:* *Chunming Zhang, University of Wisconsin-Madison;* *Zhengjun Zhang, University of Wisconsin-Madison*	E22-2002	12 Dec, 09:00 - 10:30

Latent Network Structure Learning from High Dimensional Multivariate Point Processes

Bao Cai, City University of Hong Kong

Learning the latent network structure from large scale multivariate point process data is an important task in a wide range of scientific and business applications. For instance, we might wish to estimate the neuronal functional connectivity network based on spiking (or firing) times recorded from a collection of neurons. To characterize the complex processes underlying the observed point patterns, we propose a new and flexible class of non-stationary Hawkes processes that allow both excitatory and inhibitory effects. We estimate the latent network structure using a scalable sparse least squares estimation approach. Using a novel thinning representation, we establish concentration inequalities for the first and second order statistics of the proposed Hawkes process. Such theoretical results enable us to establish the nonasymptotic error bound and the selection consistency of the estimated parameters. Furthermore, we describe a penalized least squares based statistic for testing if the background intensity is constant in time. We apply our proposed method to a neurophysiological data set that studies working memory.

Supervised Factor Modeling for High-Dimensional Linear Time Series

Guodong Li, The University of Hong Kong

Motivated by Tucker tensor decomposition, this paper imposes low-rank structures to the column and row spaces of coefficient matrices in a multivariate infinite-order vector autoregression (VAR), which leads to a supervised factor model with two factor modelings being conducted to responses and predictors simultaneously. Interestingly, the stationarity condition implies an intrinsic weak group sparsity mechanism of infinite-order VAR, and hence a rank-constrained group Lasso estimation is considered for high-dimensional linear time series. Its non-asymptotic properties are discussed thoughtfully by balancing the estimation, approximation and truncation errors. Moreover, an alternating gradient descent algorithm with hard-thresholding is designed to search for high-dimensional estimates, and its theoretical justifications, including statistical and convergence analysis, are also provided. Theoretical and computational properties of the proposed methodology are verified by simulation experiments, and the advantages over existing methods are demonstrated by analyzing US quarterly macroeconomic variables.

Two-sample tests for equal distributions in separable metric spaces: a unified semimetric-based approach

Jin-Ting Zhang, National University of Singapore

With the advancement of data collection techniques, researchers frequently encounter complex data objects within separable metric spaces across various domains. One common interest lies in determining whether two groups of complex data objects originate from the same population. This paper introduces and examines a fast and accurate unified semimetric-based approach designed to tackle this challenge. The approach exhibits broad applicability across a wide range of research areas, such as bioinformatics, audiology, environmentology, finance, and more. It effectively identifies differences between the distributions of two complex datasets, including both high-dimensional data and functional data. The asymptotic null and alternative distributions of the proposed test statistic are established.

Unlike the permutation approach, a unified, rapid and precise method to approximate the null distribution is described. Furthermore, the proposed test is shown to be root-n consistent. Numerical results are presented for illustrating the excellent performance of the proposed test in terms of size control, power, and computational cost. Additionally, the applications of the proposed test are showcased through examples involving both high-dimensional data and functional data.

Dynamic modeling via autoregressive conditional GB2 for cross-sectional maxima of financial time series data

ZhengJun Zhang, University of Wisconsin-Madison

This paper introduces the `autoregressive conditional generalized beta distribution of the second kind (AcGB2) to model the dynamic cross-sectional maxima of multivariate financial times series. The temporal dependence of the resulting univariate maxima time series is characterized by the parameter dynamics of the standard GB2 distribution, which offer versatility in approximating various distributions, including heavy-tailed distributions, through parameter adjustments. Consequently, the newly proposed AcGB2 modeling enhances flexibility in fitting real data, especially in scenarios where extreme value theory conditions are not met, as demonstrated through simulations. Real data analysis is conducted on three datasets: two with medium-sized cross-sectional dimensions (30 or fewer) and one with high dimension (100 or higher). These datasets are based on the daily negative simple-returns of 30 stocks in the Dow Jones Industrial Average, stocks in S&P 100, and stocks from 22 primary dealers, respectively. The paper establishes stationary and ergodic solutions for this new time series model under mild parameter conditions and derives the consistency, asymptotic normality, and uniqueness of the statistical conditional maximum likelihood estimators. Joint work with Ning Fan and Chunming Zhang.

Session 05	**Statistical Analysis and Inference with High Dimensional Data with Complex Structures** *Organizer/Chair: Ming-Yen Cheng, Hong Kong Baptist University*	E22-2007	12 Dec, 09:00 - 10:30

Machine Learning for Analyzing High-Dimensional Spatial Transcriptomics and Histology Image Data

Mingyao Li, University of Pennsylvania

The rapid development of spatial transcriptomics (ST) technologies has enabled the measurement of gene expression within its original tissue context. These advancements allow researchers to characterize spatial gene expression patterns, study cell-cell communications, and elucidate the spatiotemporal order of cellular development, significantly transforming our understanding of tissue functional organization. Previous studies have demonstrated a correlation between gene expression patterns and histological features, suggesting that gene expression could be predicted from histology images. However, existing methods often underutilize the rich cellular information provided by high-resolution histology. In this talk, I will present our recently developed methods designed to integrate gene expression with histology, enabling the computational reconstruction of ST data that encompasses the entire transcriptome with near-single-cell resolution. We will explore how the resulting super-resolution gene expression data can be further integrated with histology images to detect fine-grained tissue structures.

Advancing Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records

Qi Long, University of Pennsylvania

Rapid advances in technologies have enabled generation and collection of vast amounts of health data in research studies, from healthcare delivery, and from other real-world sources. While such rich data offer great promises in advancing intelligent and equitable health and medicine, they present daunting analytical challenges. One notable example is the multi-modal data from electronic health records (EHR) that are recorded at irregular time intervals with varying frequencies and include structured data such as labs and vitals, codified data such as diagnosis and procedure codes, and unstructured data such as clinical notes and pathology reports. They are typically incomplete and fraught with other data errors and biases. What's more, data gaps and errors in EHRs are often unequally distributed across patient groups: People with less access to care, often people of color or with lower socioeconomic status, tend to have more incomplete EHRs. Such data bias, if not adequately addressed, would lead to biased results and exacerbate health inequities. In this talk, I will share my research group's recent works on developing robust statistical and AI/ML methods for addressing these challenges including large language models (LLMs). I will also discuss some open questions and opportunities for future research.

A Study on Mathematical Reasoning using Functional Near-Infrared Spectroscopy

Ying Zhu, Nanyang Technological University

Mathematical reasoning, which involves making sense of mathematical ideas and concepts inherent to procedures, is essential for learning mathematics. Contemporary cognitive neuroscience has begun exploring the origins of human mathematical abilities, mainly focusing on studies of developing arithmetic skills. However, there is limited research on the origins of reasoning, one of the most complex cognitive operations. This study uses functional near-infrared spectroscopy (fNIRS) to examine brain activation changes during mathematical reasoning process among university students. Multivariate statistics and machine learning methods analyze how different functional brain systems and regions are associated with mathematical reasoning process. Investigating these dynamics could enhance our *understanding of the sources of individual differences in mathematics learning and inform effective pedagogical strategies to accelerate cognitive progress in mathematics reasoning*

Inference for possibly misspecified generalized linear models with non-polynomial dimensional nuisance parameters

Haofeng Wang, Hong Kong Baptist University

It is a routine practice in statistical modeling to first select variables and then make inference for the selected model as in stepwise regression. Such inference is made upon the assumption that the selected model is true. However, without this assumption, one would not know the validity of the inference. Similar problems also exist in high dimensional regression with regularization. To address these problems, we propose a dimension-reduced generalized likelihood ratio (DR-GLR) test for generalized linear models with non-polynomial dimensionality, based on the quasi-likelihood estimation which allows for misspecification of the conditional variance. The test has nearly oracle performance when using the correct amount of shrinkage and has robust performance against the choice of regularization parameter across a large range. We further develop an adaptive data-driven DR-GLR test and prove that with probability going to one it is an oracle GLR test. However, in ultrahigh-dimensional models the penalized estimation may produce spuriously important variables which deteriorate the performance of test. To tackle this problem, we introduce a cross-fitted DR-GLR test, which is not only free of spurious effects but robust against the choice of regularization parameter. We establish limiting distributions of the proposed tests. Their advantages are highlighted via theoretical and empirical comparisons to some competitive tests. Extensive simulations demonstrate more favorable finite sample performance of the proposed tests. An application to breast cancer data illustrates the use of our proposed methodology.

Session 09	**Recent Advances of Causal Inference** *Organizer/Chair: Weichi Wu, Tsinghua University*	E22-2009	12 Dec, 09:00 - 10:30

Data-Driven Policy Learning for a Continuous Treatment

Haitian Xie, Peking University

This paper examines policy learning for continuous treatments under unconfoundedness. The continuous-treatment scenario presents greater challenges than the discrete case because welfare estimation becomes a nonparametric issue, even with a known propensity score. This complication leads to the interaction between two nonparametric stages in policy learning: welfare estimation and policy design, and their respective tuning parameters. The bandwidth for welfare estimation must be selected simultaneously with the sieve index for policy space approximation. We contribute by considering two types of policy estimators: semi- and fully data-driven. In the semi-data-driven approach, we pre-specify the relationship between the bandwidth and the sieve index and optimize the index through structural risk minimization. In the fully data-driven approach, this relationship is determined adaptively based on data. We derive oracle inequalities for welfare for both methods, with all procedures implementable using double robustness techniques. Additionally, we identify other issues unique to the continuous treatment case, such as the theoretical properties of different penalties that were not prominent in the discrete case.

Root-n consistent semiparametric learning with high-dimensional nuisance functions under minimal sparsity

Yuhao Wang, Tsinghua University

Treatment effect estimation under unconfoundedness is a fundamental task in causal inference. In response to the challenge of analyzing high-dimensional datasets collected in substantive fields such as epidemiology, genetics, economics, and social sciences, various methods for treatment effect estimation with high-dimensional nuisance parameters (the outcome regression and the propensity score) have been developed in recent years. However, it is still unclear what is the necessary and sufficient sparsity condition on the nuisance parameters such that we can estimate the treatment effect at 1 / \sqrt{n}-rate. In this paper, we propose a new Double-Calibration strategy that corrects the estimation bias of the nuisance parameter estimates computed by regularized high-dimensional techniques and demonstrate that the corresponding Doubly-Calibrated estimator achieves 1 / \sqrt{n}-rate as long as one of the nuisance parameters is sparse with sparsity below n / \sqrt{\log p}, where p denotes the ambient dimension of the covariates, whereas the other nuisance parameter can be arbitrarily complex and completely misspecified. The Double-Calibration strategy can also be applied to settings other than treatment effect estimation, e.g. regression coefficient estimation in the presence of a diverging number of controls in a semiparametric partially linear model, and local average treatment effect estimation with instrumental variables. This is based on a joint work with Lin Liu and Xinbo Wang.

Regression and Nonparametric Adjustments for Estimating Average Treatment Effect with Network Interference

Weichi Wu, Tsinghua University

We consider the large-sample asymptotics of average treatment effect estimation under network interference within the Neyman-Rubin potential outcomes framework, assuming a setting where treatment assignments are independent of covariates and the exposure network follows a graphon model. We establish a central limit theorem for a regression-adjusted estimator, demonstrating its optimality in achieving minimal asymptotic variance within a class of linear adjustments. Additionally, we propose a novel, consistent estimator for the asymptotic variance. Furthermore, we introduce a

novel nonparametric estimator employing kernel and trimming techniques, and show its asymptotic normality, with the asymptotic variance achieving the minimum within a broader class of nonlinear adjustments. We examine the effectiveness and usefulness of our estimators through extensive simulations, and show how our results can be applied practically via a real data example. Overall, our findings confirm the promise of the regression-based estimator and highlight a path to achieving lower asymptotic variance with a kernel-and-trimming-based nonparametric estimator.

Session 12	**Dynamic Modeling of Heterogeneous Data** *Organizer/Chair: Heng Peng, Hong Kong Baptist University*	E22-2010	12 Dec, 09:00 - 10:30

A Robust Two-sample Test for High-dimensional Means

Jinfeng Xu, City University of Hong Kong

Two streams of two-sample tests for high-dimensional data have been recently studied, including the sum-of-squares-based and supreme-based tests. The former stream of tests is more powerful against dense differences in two population means, while the latter is more powerful against sparse differences. However, the level of sparsity and signal strength (i.e., magnitude of the mean differences) are often unknown in practice. It is unclear which type of tests should be applied. Motivated by this, this paper develops a robust testing procedure to provide an overall good power against a wide variety of alternative hypotheses with unknown sparsity level and varying signal strengths. The basic idea of the proposed testing procedure is to first allocate different weights onto components with varying magnitudes in a sum-of-squares-based test, and then to combine multiple weighted component tests (WCTs) to make the underlying test adaptive to different sparsity levels of the mean differences. The asymptotic properties of the proposed test are studied. Numerical comparisons demonstrate the superior performance of the proposed test across a wide spectrum of situations.

Dynamic clustering for panel data via time-varying mixture model

Youquan Pei, Shandong University

This paper introduces a time-varying mixture model to capture dynamic clustering in panel data, accommodating evolving relationships among variables and latent group structures over time. Using local maximum likelihood estimation with kernel smoothing, we derive estimators for time-varying functions and establish their asymptotic properties. A test statistic is proposed to examine the time-invariance of the mixing proportions. Simulation studies validate the model's effectiveness in identifying true cluster transitions, estimating time-varying coefficients and mixing probabilities. In an application to Environmental Kuznets Curve (EKC) data, the model reveals dynamic income-pollution relationships and changing group memberships, offering a flexible tool for analyzing complex, time-dependent data in economics and finance.

Probabilistic Principal Component Analysis with Mixture of Exponential Power Distributions

Zhenghui Feng, Harbin Institute of Technology, Shenzhen

This paper introduces the exponential power mixtures of probabilistic principal component analysis (EP-MPPCA), which serves as a flexible and robust alternative to conventional Gaussian-based mixtures of probabilistic principal component analysis (MPPCA) for data analysis. The EP-MPPCA model utilizes the exponential power distribution family, making it more adept at handling heterogeneous data distributions and outliers. Algorithms and estimation methods for the EP-MPPCA model are provided and the performance is evaluated through simulations. In real data analysis, we demonstrate how the EP-MPPCA model can be practically applied in important applications: unsupervised clustering and image data reconstruction. Specifically, we show that the EP-MPPCA model effectively handles outliers in image data, leading to improved reconstruction quality. Additionally, the model can achieve superior clustering results in an unsupervised manner.

Decorrelated forward regression for high-dimensional data analysis

Xuejun Jiang, Southern University of Science and Technology

Forward regression (FR) is a crucial methodology for automatically identifying important predictors from a large pool of potential covariates. While forward selection techniques achieve screening consistency *in contexts with moderate predictor correlation, this property gradually becomes invalid when dealing with substantially correlated variables—especially in high-dimensional datasets where strong correlations exist among predictors. This challenge is not unique to forward selection methods and is encountered by other model selection approaches as well. To address these challenges, we introduce a novel decorrelated forward (DF) selection framework for generalized mean regression models, including prevalent models, such as linear, logistic, Poisson, and quasi likelihood. The DF selection framework stands out because of its ability to convert generalized mean regression models into linear ones, thus providing a clear interpretation of the forward selection process. It also offers a closed-form expression for forward iteration, to improve practical applicability and efficiency. Theoretically, we establish the screening consistency of DF selection and determine the upper bound of the selected submodel's size. To reduce computa-tional burden, we develop a thresholding DF algorithm that provides a stopping rule for the forward-searching process. Simulations and real data applications show the outstanding performance of our method compared with that of some existing model selection methods.*

Session 14	**Random Matrix and It's Application in Finance** *Organizer/Chair: Shurong Zheng, Northeast Normal University; Yongchang Hui, Xi'an Jiaotong University*	E22-2011	12 Dec, 09:00 - 10:30

Pragmatic attitude to large-scale Markowitz's portfolio optimization and factor-augmented derating

Yongchang Hui, Xi'an Jiaotong University

In this paper, we propose a Factor-Augmented Derating (FAD) method for large-scale mean–variance portfolio optimization to further overcome the overprediction phenomenon pointed by Bai et al., (2009). They found out the optimal return obtained by plug-in method was consistently higher than the theoretical optimal return and proposed a bootstrap de-rated optimal return instead based on random matrix theory. Incorporating the widely recognized fact in empirical finance studies that high-dimensional stock returns often conform to factor models, we replace the estimator of the precision matrix with a low-rank estimator in the plug-in optimal return, and further derate it using the correction parameter derived from Bai et al., (2009). We establish theories to verify why the FAD method can more effectively avoid overprediction. In our simulation, we find that derating is requisite and our FAD optimal return is the closest to the theoretical optimal return comparing to plug-in, bootstrap-derated and factor-based optimal returns in high-dimensional situations. We also find that the FAD optimal return is the most credible in our empirical studies on portfolio allocation among 200 component stocks of S&P500. Backtesting results clearly show that the discrepancy of "high expectation-low realization" can be best reduced by using the FAD method, though no real returns can achieve the anticipated optimal returns. More surprisingly, FAD method yields the highest real returns, even with low optimal returns at the decision-making stage.

High dimensional homogeneity test under an unsupervised context

Yiming Liu, Ji'nan University

In this paper, we introduce a novel homogeneity test for high-dimensional data within the framework of unsupervised learning. Utilizing the spiked singular value of the data matrix under both null and alternative hypotheses, we propose a new singular-vector-based statistic designed to distinguish between two distinct types of signals. To establish the limiting distribution of the proposed statistic, we first derive the empirical measure of the normalized singular vector associated with the spiked singular value, applicable under both the null and alternative conditions. Furthermore, our simulation studies, along with analyses of real data from breast cancer and glioblastoma gene expression datasets, underscore the efficacy and applicability of the proposed methodology.

A Framework for Statistical Inference via Randomized Algorithms

Zhixiang Zhang, University of Macau

Randomized algorithms, such as randomized sketching or stochastic optimization, are a promising approach to ease the computational burden in analyzing large datasets. However, randomized algorithms also produce non-deterministic outputs, leading to the problem of evaluating their accuracy. In this paper, we develop a statistical inference framework for quantifying the uncertainty of the outputs of randomized algorithms. Our key conclusion is that one can perform statistical inference for the target of a sequence of randomized algorithms as long as in the limit their outputs fluctuate around the target according to any (possibly unknown) probability distribution. In this setting, we develop appropriate statistical inference methods---sub-randomization, multi-run plug-in and multi-run aggregation---by estimating the unknown parameters of the limiting distribution either using multiple runs of the randomized algorithm, or by tailored estimation. As illustrations, we develop methods for statistical inference for least squares parameters via random sketching (sketch-and-solve, partial and iterative

sketching), by characterizing their limiting distribution in a possibly growing dimensional case. Moreover, we also apply our inference framework to stochastic optimization, including for stochastic gradient descent and stochastic optimization with momentum. The results are supported via a broad range of simulations.

Session 27	**Advanced Techniques in Design of Experiments** *Organizer/Chair:* *Lianjie Shu, University of Macau;* *Yongxiang Li, Shanghai Jiao Tong University*	E22- 2013	12 Dec, 09:00 - 10:30

Probabilistic battery health evaluation via deep Gaussian mixture density network

Zhelin Huang, Shenzhen University

Accurate state-of-health estimation of lithium-ion batteries is crucial for ensuring the safety and reliability of electric vehicles. A data-driven method is proposed for realizing accurate and robust state-of-health estimation with uncertainty measures. First, a set of battery health features are proposed based on the original and differential capacity-voltage and temperature-voltage curves for depicting the multitimescale battery ageing behavior. Next, a deep learning-assisted Gaussian mixture density network is developed to fuse the proposed features and generate the conditional probability distribution of battery state-of-health. Finally, a comprehensive computational study is conducted based on three datasets, which contain batteries of different chemistries and operated under different conditions. Results verify that the proposed method can generate an accurate and robust state-of-health estimation as well as provide the probability measure.

Uniform designs of experiments with mixtures under the criterion mean L1-distance and a new approach to Scheffé-type designs

Yaping Wang, East China Normal University

Mixture experiments analyze how changes in component proportions impact the response variable within the experimental region of a simplex. This paper introduces a new criterion, named the mean L1-distance (ML1D) criterion, for constructing uniform designs in mixture experiments. This criterion allows flexibility in point size and showcases a more uniform pattern within the experimental region. We also explore the optimal Scheffé-type simplex-lattice designs under the ML1D criterion. An interesting discovery is that the uniform mixture designs and the optimal Scheffé-type simplex-lattice designs are connected. For a two-component mixture design, these two types of designs are proven to be equivalent. For more than two-component mixture designs, numerical equivalences between the two designs are observed. These findings strengthen the rationale for users to adopt these designs in mixture experiments for modeling and prediction.

Multi-fidelity Gaussian process modeling with boundary information

Matthias HY Tan, City University of Hong Kong

Time-consuming bi-fidelity simulations with a high-fidelity (HF) simulator and a lowfidelity (LF) simulator, where the HF simulator contains a vector of inputs not shared with the LF simulator, called the augmented input, arise frequently in practice. For such simulations, it is frequently known a priori that when the augmented input converges to any value in a subset of the boundary of its domain, the HF simulator output converges to the LF simulator output. This is a form of boundary information, i.e., prior information on the output of a simulator at the boundary of the domain of the simulator's inputs. To reduce simulation time, the standard autoregressive Gaussian process (GP) emulator can be constructed to approximate and replace the HF and LF simulators. However, this emulator does not satisfy boundary information. In this talk, I will present a solution to the problem of constructing a bi-fidelity GP emulator that satisfies the form of boundary information just mentioned. The proposed emulator, called the boundary modified autoregressive GP (BMAGP) emulator, is shown to outperform the standard autoregressive GP emulator with some examples.

Penalized Additive Gaussian Process for Screening and Optimization of Quantitative and Qualitative Factors in Black-Box Systems

Yongxiang Li, Shanghai Jiao Tong University

Variable screening and Bayesian optimization of both quantitative and qualitative (QQ) factors are critical in various applications where evaluating black-box systems is resource-intensive or time-consuming. Traditional sensitivity analysis lacks a unified framework for simultaneously screening important QQ factors and struggles to screen important levels of a qualitative factor (factor levels). To address this, we introduce a penalized additive Gaussian process (PAGP) model, featuring an interpretable additive (IA) covariance function for QQ factors. This allows for sparsity penalties that enable the identification of critical factor levels. A gradient-informed optimization approach using derivative information is proposed to accelerate PAGP modeling, and a tailored ADMM is proposed to optimize the L1-regularized likelihood. Then, this study proposes factor level screening utilizing sparse regularization and quantitative factor screening leveraging the Shapley value. Finally, Bayesian optimization is introduced to PAGP for optimizing black-box systems with QQ factors, which also provides an interpretable importance quantification of factor levels during optimization. PAGP distinguishes itself by enabling sparse regularization and efficient screening of factor levels. Simulation studies validate the performance of PAGP, and it is also applied to the design of paper pilots and neural networks.

Grouped orthogonal arrays for computer experiments

Wenlong Li, Beijing Jiaotong University

We propose methods for constructing a new type of space-filling design known as a grouped orthogonal array, designed to accommodate natural input grouping in computer experiments. This design achieves space-filling properties across all inputs and exhibit stronger space-filling properties within inputs within the same groups compared to those from different groups. Using combinatorial orthogonality as a guiding space-filling criterion, this method generates a class of grouped orthogonal arrays, consisting of special strength-two orthogonal arrays with columns partitioned into groups of strength-three orthogonal arrays. The proposed methods are straightforward to implement and can handle a large number of factors. Examples are provided to illustrate the methods, and comparisons with other space-filling designs demonstrate the importance of exploring the grouping structure in our proposed design. Simulations are presented to illustrate the effectiveness of the proposed designs for emulating computer models.

Session 31	**Optimization and Statistics for Large Language Models** *Organizer/Chair:* *Jiancong Xiao, University of Pennsylvania;* *Ruochen Jin, East China Normal University*	E22-2014	12 Dec, 09:00 - 10:30

Divergent LLM Adoption and Heterogeneous Convergence Paths in Research Writing

Wu Zhu, Tsinghua University

The emergence of Large Language Models (LLMs, e.g., ChatGPT) is believed to revolutionize writing. We investigate the impact of generative-AI-assisted revisions on academic writing, focusing on researchers' heterogeneous usage and their convergence in writing. Using a dataset of over 627,000 academic papers from arXiv, we develop a framework for identifying ChatGPT-revised articles by training a set of prompt- and discipline-specific language models ourselves. We first document the widespread usage of GPT, with significant heterogeneity across disciplines, gender, ethnicity, and academic experience, as well as a rapid evolution in writing style over time. Moreover, our analysis reveals that LLM usage significantly improves writing outcomes in terms of clarity, conciseness, adherence to formal writing rules, etc., and the improvements depend on nuanced types of usage. Finally, a difference-in-difference analysis suggests that while the birth of LLMs leads to a convergence in academic writing, adopters, males, non-native speakers, and junior researchers adjust their writing style the most to resemble that of more experienced scholars. Our findings raise concerns about potential homogenization of academic expression and the unequal benefits researchers derive from new technologies.

Magnetic Mirror Descent Self-play Preference Optimization

Mingzhi Wang, Peking University

Standard Reinforcement Learning from Human Feedback (RLHF) methods mainly optimize preferences through the Bradley-Terry (BT) reward model, which may misalign with natural human preferences due to the strong transitivity assumption. Recent work has reframed the preference learning problem as a two-player constant-sum game, aiming to learn policies that better reflect human preferences by finding the Nash equilibrium (NE) of this game. However, existing methods under this framework either guarantee only average-iterate convergence or rely on strong first-order approximation assumptions. In this paper, we propose Mirror Descent Self-play Preference Optimization (MDSPO), a novel approach based on Magnetic Mirror Descent (MMD). By introducing an additional magnetic term, MDSPO achieves linear convergence rate to the NE of the regularized game. Furthermore, we establish theoretical guarantees for the convergence of our algorithm to the NE of the original game by periodically updating the reference policy. This approach effectively guarantees that the final policy accurately reflects the true human preferences. To ensure our algorithm is both theoretically sound and practically viable, we provide a simple yet effective implementation that adapts the theoretical insights to the RLHF setting. We demonstrate its effectiveness on a variety of benchmarks.

Fine-Tuning Attention Modules Only: Enhancing Weight Disentanglement in Task Arithmetic

Ruochen Jin, East China Normal University

For the past several years, task arithmetic has gained increasing attention. This approach edits pre-trained models directly in weight space by combining the fine-tuned weights of various tasks into a unified model. Its efficiency and cost-effectiveness stem from its training-free combination, contrasting with traditional methods that require model training on large datasets for multiple tasks. However, applying such a unified model to individual tasks can lead to interference from other tasks (lack of weight disentanglement). To address this issue, Neural Tangent Kernel (NTK) linearization has been employed to leverage a "kernel behavior", facilitating weight disentanglement and mitigating adverse effects from *unrelated tasks. Despite its benefits, NTK linearization presents drawbacks, including doubled training costs, as well as reduced performance of individual models. To tackle this problem, we propose a simple yet effective and efficient method that is to finetune the attention modules only in the Transformer. Our study reveals that the attention modules exhibit kernel behavior, and fine-tuning the attention modules only significantly improves weight disentanglement. To further understand how our method improves the weight disentanglement of task arithmetic, we present a comprehensive study of task arithmetic by differentiating the role of the representation module and task-specific module. In particular, we find that the representation module plays an important role in improving weight disentanglement whereas the task-specific modules such as the classification heads can degenerate the weight disentanglement performance.*

On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization

Jiancong Xiao, University of Pennsyivania

Accurately aligning large language models (LLMs) with human preferences is crucial for informing fair, economically sound, and statistically efficient decision-making processes. However, we argue that reinforcement learning from human feedback (RLHF)—the predominant approach for aligning LLMs with human preferences through a reward model—suffers from an inherent algorithmic bias due to its Kullback–Leibler-based regularization in optimization. In extreme cases, this bias could lead to a phenomenon we term preference collapse, where minority preferences are virtually disregarded. To mitigate this algorithmic bias, we introduce preference matching (PM) RLHF, a novel approach that provably aligns LLMs with the preference distribution of the reward model under the Bradley–Terry–Luce/Plackett–Luce model. Central to our approach is a PM regularizer that takes the form of the negative logarithm of the LLM's policy probability distribution over responses, which helps the LLM balance response diversification and reward maximization. Notably, we obtain this regularizer by solving an ordinary differential equation that is necessary for the PM property. For practical implementation, we introduce a conditional variant of PM RLHF that is tailored to natural language generation. Finally, we empirically validate the effectiveness of conditional PM RLHF through experiments on the OPT-1.3B and Llama-2-7B models, demonstrating a 29% to 41% improvement in alignment with human preferences, as measured by a certain metric, compared to standard RLHF.

Session 33	**Time Series Econometrics** *Organizer/Chair: Degui Li, University of Macau*	E22-2015	12 Dec, 09:00 - 10:30

Multiple-use calibration for all future values: simultaneous tolerance regions for multivariate regression

Yang Han, University of Manchester

Multiple-use calibration for all future values plays a valuable role in many fields, including industry, health and medical research. Simultaneous tolerance regions (STRs) can be used for this purpose in multivariate regression; however, no construction methods are currently available in the literature. This work first fills the gap by proposing methods for constructing STRs, providing a solution to multiple-use calibration problems. We also develop weighted simultaneous tolerance regions (WSTRs), showing that confidence sets based on WSTRs can exactly satisfy the key property for multiple-use calibration. Through the lens of WSTRs, we address the misconception that the confidence sets derived from pointwise tolerance regions can guarantee the key property. Furthermore, we propose several methods for computing the critical constants for STRs and WSTRs. Simulation studies are conducted for comparing these methods in terms of accuracy and computational cost. Real-data examples are provided for illustration.

Multimodality in Meta-Learning and Opportunities in Tourism

Albert Li, Wynn Palace

Meta-learning has gained wide popularity as a training framework that is more data-efficient than traditional machine-learning methods. However, its generalization ability in complex task distributions, such as multimodal tasks, has not been thoroughly studied. Recently, some studies on multimodality-based meta-learning have emerged. This talk provides an overview of the multimodality-based meta-learning landscape in terms of the methodologies. We will also discuss potential applications and opportunities in tourism.

Large-Scale Curve Time Series with Common Stochastic Trends

Degui Li, University of Macau

In this paper, we study high-dimensional curve time series with common stochastic trends. We adopt a dual functional factor model structure with a high-dimensional factor model for the observed curve time series and a low-dimensional factor model for the latent curves with common trends. A functional PCA technique is applied to estimate the common stochastic trends and functional factor loadings. Under some regularity conditions, we derive the mean square convergence and limit distribution theory for the developed estimates, allowing the dimension and sample size to jointly diverge to infinity. We also propose an easy-to-implement criterion to consistently select the number of common stochastic trends and further discuss the model estimation when the nonstationary factors are cointegrated. Extensive Monte-Carlo simulation studies and two empirical applications to large-scale temperature curves in Australia and log-price curves of the S\&P stocks, respectively, are conducted to illustrate the finite-sample performance of the developed methodology.

Spurious Quantile Regressions and Variable Selection in Quantile Cointegrations

Siwei Wang, Hunan University

Quantile regression is an effective tool in modeling conditional distribution of economic and financial time series. This paper first investigates the spurious quantile regression phenomenon involving processes moderately deviated from a unit root (PMDURs) through numerical experiments. The main findings include large quantile correlation coefficient estimates, divergent significance test statistics, *high R2 and highly autocorrelated residuals indicated by very low Durbin-Watson statistics, complementing those discovered in spurious mean regressions (Lin and Tu, 2020). Unlike the recently proposed balanced regression (Ren et al., 2019; Lin and Tu, 2020) for the mean regression, this paper pioneers the use of the quantile partial correlation (Li et al., 2015, QPC) as a simple-to-implement robust inference measure for spurious quantile regressions.*

This is benefited from our finding that the induced QPC estimator is asymptotically normal and free of nuisance parameters. For the correlated but non-cointegrated quantile regressions and quantile cointegrations, the QPC estimator converges to a nonzero value in probability. Consequently, the proposed inference procedure can serve to perform variable selection in quantile regressions/cointegrations with PMDURs. Finally, the finite sample properties of the robust method are demonstrated through both Monte Carlo and real data examples.

Session 42	**High Dimensional Statistics Inference** *Organizer/Chair: Guangming Pan, Nanyang Technological University*	E22-2017	12 Dec, 09:00 - 10:30

Vast Portfolio Selection with Constrained Dantzig-type Estimator

Patrick Pun, Nanyang Technological University

Asset selection problems, especially sparse portfolio construction, are getting uprising attention in recent years, due to the vast increase in the number of available assets to choose from, paired with comparably limited yet noisy information in the market. Moreover, the investment activities are often restricted by various constraints, e.g. budget constraint. In this paper, we develop constrained Dantzig-type estimator (CDE) for sparse learning problems with equality constraints, such as sparse portfolio construction. We show that CDE is able to produce an estimate to the oracle solution and counter the curse of dimensionality, while we derive its non-asymptotic statistical error bounds under l_1 and l_2 norms. Compared to the constrained Lasso, CDE has significant computational advantage as we can obtain CDE via the solution to a linear program. Moreover, CDE is extremely versatile and widely applicable. Through extensive simulations and empirical studies, we show that sparse portfolios constructed using CDE have superior out-of-sample performance compared to various benchmark portfolios, including the equally weighted portfolio.

Penalized Principal Component Analysis for Large-dimension Factor Model with Group Pursuit

Yiming Wang, Shandong University

This paper investigates the intrinsic group structures within the framework of large-dimensional approximate factor models, which portrays homogeneous effects of the common factors on the individuals that fall into the same group. To this end, we propose a fusion Penalized Principal Component Analysis (PPCA) method and derive a closed-form solution for the ℓ_2-norm optimization problem. We also show the asymptotic properties of our proposed PPCA estimates. With the PPCA estimates as an initialization, we identify the unknown group structure by a combination of the agglomerative hierarchical clustering algorithm and an information criterion. Then the factor loadings and factor scores are re-estimated conditional on the identified latent groups. Under some regularity conditions, we establish the consistency of the membership estimators as well as that of the group number estimator derived from the information criterion. Theoretically, we show that the post-clustering estimators for the factor loadings with group pursuit achieve efficiency gains compared to the estimators by traditional PCA method. Thorough numerical studies validate the established theory and a real financial example illustrates the practical usefulness of the proposed method.

Liberating dimension and spectral norm: A universal approach to spectral properties of sample covariance matrices

Yanqing Yin, Chongqing University

In this paper, our primary objective is to elucidate a guiding principle that governs the spectral properties of the sample covariance matrix. This principle exhibits a harmonious behavior across various limiting frameworks, eliminating the necessity for constraints on the rates of dimension p and sample size n as long as both tend to infinity. We achieve this by employing a well-suited normalization technique on the original sample covariance matrix. Subsequently, we establish a robust central limit theorem for linear spectral statistics within this expansive framework, extending the Bai-Silverstein theorem (Ann Probab 32(1):553-605, 2004). This accomplishment effec- tively eliminates the need for a bounded spectral norm on the population covariance matrix and relaxes constraints on the rates of dimension p and sample size n. As a result, our findings significantly broaden the applicability of these results in the realm of high-dimensional statistics. To demonstrate the potency of our established results, we provide *an illustrative example involving the test for covariance structure under high dimensionality. This illustrative example extends the findings in the work of Ledoit and Wolf (Ann Stat 30: 1081–1102, 2002) and Qiu, Li, and Yao (Ann Stat 51: 1427–1451, 2023) by liberating both p and n. Extensive numerical analyses are conducted to thoroughly investigate the robustness of our theoretical findings.*

Hypothesis tests for high-dimensional MA and white noise

Chi Yao, Nanyang Technological University

Hypothesis testing for observations in time series has gained significant attention, particularly for high-dimensional complex-valued linear processes. In this context, we propose a test statistic based on the trace of the product of the sample autocovariance matrix and its transpose at adjacent time lags. As both the dimension p and the sample size n diverge to infinity, we relax the typical requirement that $p/n \to c > 0$, instead requiring only $p = O(n^{2 - \epsilon})$ for any $\epsilon > 0$. The test statistic’s asymptotic normality is established when the lag τ exceeds the model order q. This approach circumvents the need for estimating the coefficient matrix and the fourth moments of the error term. Based on these properties, we develop hypothesis tests for high-dimensional complex-valued linear models. When $q = 0$, the null hypothesis tests for white noise, with alternatives including VMA, VAR, and VARMA processes. For $q > 0$, it tests for VMA models, with alternatives including white noise, VAR, and VARMA.

Session 44	**Recent Advances in Large Language Models and Large-scale Data Analysis** *Organizer/Chair:* *Xinyuan Song, The Chinese University of Hong Kong;* *Zhixiang Lin, The Chinese University of Hong Kong*	E22-2018	12 Dec, 09:00 - 10:30

De Novo Functional Protein Design: Overcoming Data Scarcity with Regeneration and Large Sequence Models

Jian Huang, The Hong Kong Polytechnic University

Proteins are essential components of all living organisms and play a critical role in cellular survival. They have a broad range of applications, from clinical treatments to material engineering. This versatility has spurred the development of protein design, with amino acid sequence design being a crucial step in the process. Recent advancements in deep generative models have shown promise for protein sequence design. However, the scarcity of functional protein sequence data for certain types can hinder the training of these models, which often require large datasets. To address this challenge, we propose a hierarchical model named ProteinRG that can generate functional protein sequences using relatively small datasets. ProteinRG begins by generating a representation of a protein sequence, leveraging existing large protein sequence models, before producing a functional protein sequence. We have tested our model on various functional protein sequences and evaluated the results from three perspectives: multiple sequence alignment, t-SNE distribution analysis, and 3D structure prediction. The findings indicate that our generated protein sequences maintain both similarity to the original sequences and consistency with the desired functions. Moreover, our model demonstrates superior performance compared to other generative models for protein sequence generation. This is joint work with Chenyu Ren and Daihai He.

An alternative measure for quantifying the heterogeneity in meta-analysis

Tiejun Tong, Hong Kong Baptist University

Quantifying the heterogeneity is an important issue in meta-analysis, and among the existing measures, the I^2 statistic is most commonly used. In this paper, we first illustrate with a simple example that the I^2 statistic is heavily dependent on the study sample sizes, mainly because it is used to quantify the heterogeneity between the observed effect sizes. To reduce the influence of sample sizes, we introduce an alternative measure that aims to directly measure the heterogeneity between the study populations involved in the meta-analysis. We further propose a new estimator, namely the I_A^2 statistic, to estimate the newly defined measure of heterogeneity. For practical implementation, the exact formulas of the I_A^2 statistic are also derived under two common scenarios with the effect size as the mean difference (MD) or the standardized mean difference (SMD). Simulations and real data analysis demonstrate that the I_A^2 statistic provides an asymptotically unbiased estimator for the absolute heterogeneity between the study populations, and it is also independent of the study sample sizes as expected. To conclude, our newly defined I_A^2 statistic can be used as a supplemental measure of heterogeneity to monitor the situations where the study effect sizes are indeed similar with little biological difference. In such scenario, the fixed-effect model can be appropriate; nevertheless, when the sample sizes are sufficiently large, the I^2 statistic may still increase to 1 and subsequently suggest the random-effects model for metaanalysis.

Testing for Change-points in Heavy-tailed Time Series – A Winsorized CUSUM Approach

Shiqing Ling, Hong Kong University of Science and Technology

It is well-known how to detect the change-point in heavy-tailed time series is an open problem since the traditional tests may not have a power. This article proposes a winsorized CUSUM approach to solve this problem. We begin by investigating the winsorized CUSUM process and deriving the limiting *distributions of the Kolmogorov-Smirnov test and the Self-normalized test under the null hypothesis. Under the alternative hypothesis, we firstly uncover the behavior of change-point magnitude after the winsorized data and show that our tests have a power approaching to 1 as the sample size n ! 1. We then extend the winsorizing technique to tests for multiple change-points without the prior information on the number of actual change points. Our framework is quite general and its assumption is very weak. This enables the application of our tests to both linear time series and nonlinear time series, such as TAR and G-GARCH processes. The empirical results illustrate the effectiveness of our proposed procedures for change-point detection. (This is a joint work with She Rui and Da Linlin)*

Tracking structural changes in dynamic heterogeneous networks

Junhui Wang, The Chinese University of Hong Kong

Dynamic networks consist of a sequence of time-varying heterogeneous networks, and it is of great importance to detect the structural changes. Most existing methods focus on detecting abrupt network changes, necessitating the assumption that the underlying network probability matrix remains constant between adjacent change points. This assumption can be overly strict in many real-life scenarios due to their versatile network dynamics. In this talk, we introduce a new subspace tracking method to detect network structural changes in dynamic networks, whose network connection probabilities may still undergo continuous changes. Particularly, two new detection statistics are proposed to jointly detect the network structural changes, followed by a carefully refined detection procedure. Theoretically, we show that the proposed method is asymptotically consistent in terms of detecting the network structural changes, and also establish the impossibility region in a minimax fashion. The advantage of the proposed method is supported by extensive numerical experiments on both synthetic networks and a series of UK politician social networks.

Session 49	**Navigating Digital Landscapes: Insights from Platform Economy** *Organizer/Chair: Yingpeng Zhu, University of Macau*	E22-G008	12 Dec, 09:00 - 10:30

The Effects of Copyrights on Music Creation: Evidence from the U.S. Music Modernization Act

Jiaxin Lei, The Hong Kong University of Science and Technology

Copyright laws aim to balance two competing priorities: financially motivating creators and guaranteeing public access. The digital transformation has prompted a reassessment of how to balance these two objectives in copyright protection. This paper aims to provide the first systematic evidence on copyright protection in the digital age by examining the impact of stricter copyright laws on music creation, using the Music Modernization Act (MMA) in the US as a natural experiment. We conduct a difference-in-differences analysis with US singers as the treatment group and singers in other English-speaking and non-English-speaking countries as the control group. Our results show that the implementation of the MMA, a more stringent copyright law, leads to a significant decrease in music creation, particularly among less popular and niche-type musicians. We provide evidence that the MMA results in higher costs for streaming service providers and limits public access to music, which in turn reduces the visibility and profitability of musicians and undermines their motivation to create new pieces. Our study highlights the need for policymakers to carefully evaluate the impact of copyright protection policies on content creation in the digital age.

A Smart Solution to Rush Hour Traffic Congestion: Effects of Dockless Bike-Sharing Entry on Ride-Sharing

Juan Qin, University of Science and Technology of China

Urban planners, policymakers, transportation engineers, and scholars seek innovative and sustainable solutions to rush-hour traffic congestion. This paper examines whether dockless bike-sharing systems can offer a technology-enabled solution to urban rush-hour congestion problems. We leverage a unique natural experiment setting provided by the launch of the first dockless bike-sharing service in an urban city. We employ a difference-in-differences framework and detailed ride-sharing trajectory data to examine changes in the Travel Time Index (TTI), a comprehensive spatial-temporal measure of congestion. We find that the introduction of dockless bike-sharing significantly decreases rush-hour traffic congestion. Furthermore, we examine the heterogeneity in the effect across multiple key temporal and spatial dimensions. The dockless bike-sharing system has a stronger effect in reducing congestion during commute-oriented weekday rush hours compared to leisure-oriented weekend rush hours, which highlights the system's effectiveness in alleviating commute-oriented congestion. In addition, locations with subway stations experience a greater reduction in congestion, providing empirical evidence that dockless bikes can provide effective first-mile/last-mile connections for subways. Conversely, locations with bus stations do not experience a similar reduction in congestion, suggesting that the lack of appropriate bike regulations and limited space around bus stations may lead to congestion if dockless bikes are not parked properly. We document additional important spatial variations in the congestion reduction effect. Congestion relief is more prominent in city center areas, suggesting that policymakers should consider different policies for dockless bikes depending on the centrality of urban locations. In residential areas, median-priced residential neighborhoods experience the greatest reduction in congestion, which indicates that socioeconomic factors affect individuals' dockless bike usage. Additionally, the congestion reduction effect is most pronounced on arterial and major roads, which typically experience heavy traffic during rush hours. Overall, we find that while dockless bike-sharing systems offer a promising smart solution to congestion, the effect varies across locations and transportation infrastructure, underscoring that carefully designed context-specific policies and regulations are necessary to enhance the benefits of bike-sharing services. Our findings provide *valuable insights for urban planners and policymakers in crafting effective strategies for dockless bike deployment, regulation, and usage.*

Status Goods Consumption and Gamification in Digital Markets

Lin Qiu, Southern University of Science and Technology

The success of status goods such as licensed collectibles and fashion products hinges on their ability to convey owner's status to other consumers. In digital settings, however, the absence of reliable signals of ownership can diminish these goods' ability to serve as status symbols, potentially reducing demand. To address this challenge, we advocate for using gamification to disclose consumers' purchase information. We first exam-ine how different levels of gamified information disclosure influence consumers' status goods purchases by exploiting two field designs on an e-commerce platform for status products. We then develop a theoretical model to study firms' optimal strategies for disclosing consumer purchase information. Our findings suggest that purchase information disclosure via carefully crafted gamification designs can increase the purchase of status products. Nevertheless, impacts vary by product type: a higher degree of disclosure consistently leads to increased purchases of high-priced status goods, but it might deter purchases of low-priced status products. Furthermore, the effectiveness of gamified information disclosure differs across consumer segments and is more pronounced for status-sensitive consumers. This research contributes to understanding the relationship between gamification and status goods consumption, providing valuable insights for marketers seeking to leverage gamification designs to promote status goods in online markets.

Advertising Selling, Information Sharing and Selling Format in Online Retailing

Jianyue Wang, The Hong Kong University of Science and Technology

This paper investigates the role of advertising selling in online retailing and how it impacts the selling format, information sharing strategies and firms' profits. We develop a game-theoretic modeling framework to investigate the equilibrium selling format and firms' profits in a supply chain that consists of a seller and an online platform. We solve for the equilibrium in several scenarios depending on whether or not the platform offers advertising selling and/or information sharing services to the seller. By comparing the equilibrium results, we examine the impact of advertising selling on the selling format, information sharing strategies and firms' profits. Our analysis reveals two novel effects of advertising selling, namely the cost sharing flexibility effect and wholesale price effect. The first effect increases the supply chain efficiency under the agency selling format whereas the second effect decreases it under the reselling format. The platform chooses the agency selling format and shares information under a broader set of conditions if he sells advertising than if he does not sell it. Advertising selling and information sharing services are complementary in the sense that either one of these services benefits the platform under a broader set of conditions if the other service is available than if it is not available. Our study provides useful insights to managers who make selling format, advertising selling and information sharing decisions in online retailing. It provides an explanation about how the growth of the agency selling format could be driven by the trend that more online retailers are offering advertising selling and information sharing services.

Unmasking the Ripple Effects: Platform Endorsement, Seller Reputation, and the Spillover of Online Review Manipulation

Le Wang, City University of Hong Kong

The direct consequences of online review manipulation have been well-documented, yet the extent to which such actions impact competitors and the mechanisms through which these spillover effects manifest remain largely understudied. To advance this line of research, this study delves into how the manipulation of reviews by an online seller affects the sales of its competitors, accounting for the interplay between platform endorsement and seller reputation. Drawing on trust transfer theory and the accessibility–diagnosticity framework, we propose two distinct pathways through which online review manipulation influences competitor sales. Our empirical analysis, based on over one million observations from two leading online reservation platforms in China, supplemented by two pre-registered online experiments, reveals three key findings: (1) for platform-endorsed sellers, review manipulation triggers a negative competitive effect, leading to a decrease in their competitors' sales; (2) for reputable sellers without platform endorsement, review manipulation induces a contagion effect, hindering their competitors' sales; and (3) for less reputable sellers without platform endorsement, review manipulation leads to a positive competitive effect and boosts their competitors' sales. These findings remain robust across various tests, including alternative measures of variables, different model specifications, tests for reverse spillover effects, alternative thresholds for reputation, and ruling out alternative explanations. This study contributes to the literature by reconciling contradictory predictions about how review manipulation affects competitors by considering both seller reputation and platform endorsement. It offers actionable implications for marketers and e-platforms to curb unethical competitive practices, such as review manipulation, by leveraging the power of online reputation systems and platform endorsement to effectively mitigate such adverse effects.