Invited Sessions IV - 12 December 2024, 11:00 – 12:30

Session 01	**Recent Development on Micro/Macro Data Analyses** *Organizer/Chair: Bin Peng, Monash University*	E22-2002	12 Dec, 11:00 - 12:30

Detecting Spurious Factor Models

Bo Zhang, University of Science and Technology of China

We propose a new testing for detecting spurious factor models based on random matrix theory and power law. We apply it on some interesting real data sets and show the risk of spurious factors.

A Robust Residual-Based Test for Structural Changes in Factor Models

Bin Peng, Monash University

In this paper, we propose an easy-to-implement residual-based specification testing procedure for detecting structural changes in factor models, which is powerful against both smooth and abrupt structural changes with unknown break dates. The proposed test is robust against the over-specified number of factors, and serially and cross-sectionally correlated error processes. A new central limit theorem is given for the quadratic forms of panel data with dependence over both dimensions, thereby filling a gap in the literature. We establish the asymptotic properties of the proposed test statistic, and accordingly develop a simulation-based scheme to select critical value in order to improve finite sample performance. Through extensive simulations and a real-world application, we confirm our theoretical results and demonstrate that the proposed test exhibits desirable size and power in practice.

Sparse Heteroskedastic PCA in High Dimensions

Zhao Ren, University of Pittsburgh

Principal Component Analysis (PCA) is a widely adopted multivariate statistical technique, renowned for its versatility across numerous disciplines. To address the challenges posed by high-dimensional and heteroskedastic data, we consider a general framework in high dimensions, the generalized spiked covariance model with sparse loadings. We propose a novel algorithm called \emph{SparseHPCA} that leverages the orthogonal iteration method, noise-adaptive thresholding, and diagonal imputation techniques. The proposed procedure is computationally feasible and fully data-driven without prior knowledge about the noise levels and the sparsity of the loading matrix. Theoretical analysis shows that SparseHPCA achieves minimax optimal convergence rates. By applying SparseHPCA, we further investigate the Sparse SVD problem in the presence of heteroskedastic noise. The effectiveness of the proposed method is revealed through experiments on both simulated and two real datasets.

Estimating short-term capacity: Theory and an application to invasive coronary angiography for acute myocardial infarction

Ou Yang, The University of Melbourne

This paper develops a new theoretical model to illustrate how the optimal capacity level should be chosen by the hospital given their knowledge of uncertain patient demand. Since the capacity level chosen by the hospital is often not observed and has to be estimated by researchers in practice, a novel non-parametric model is employed to estimate the short-term capacity chosen by the hospital. In an application, the role of short-term capacity in the case of treatment for acute myocardial infarction (AMI) is examined. Due to its urgency, patients suffering from AMI are often sent to the nearest hospital. The short-term capacity of the hospital in delivering care is therefore paramount for patient well-being and survival. Our results indicate significant short-term capacity effects and suggest that the effects are higher for ST-elevation myocardial infarctions (STEMI) patients than Non-ST-elevation myocardial

infarctions (NSTEMI) patients, reflecting the greater importance of short-term capacity due to the urgency of STEMI.

Session 07	**Statistical Analysis for Complex and High-dimensional Data** *Organizer/Chair: Qianqian Zhu, Shanghai University of Finance and Economics*	E22-2007	12 Dec, 11:00 - 12:30

Estimating spot volatility under infinite variation jumps with dependent market microstructure noise

Qiang Liu, Shanghai University of Finance and Economics

Jumps and market microstructure noise are stylized features of high-frequency financial data. It is well known that they introduce bias in the estimation of volatility (including integrated and spot volatilities) of assets, and many methods have been proposed to deal with this problem. When the jumps are intensive with infinite variation, the efficient estimation of spot volatility under serially dependent noise is not available and is thus in need. For this purpose, we propose a novel estimator of spot volatility with a hybrid use of the pre-averaging technique and the empirical characteristic function. Under mild assumptions, the results of consistency and asymptotic normality of our estimator are established. Furthermore, we show that our estimator achieves an almost efficient convergence rate with optimal variance when the jumps are either less active or active with symmetric structure. Simulation studies verify our theoretical conclusions. We apply our proposed estimator to empirical analyses, such as estimating the weekly volatility curve using second-by-second transaction price data.

Vector autoregressive models with common response and predictor factors

Di Wang, Shanghai Jiao Tong University

The reduced-rank vector autoregressive (VAR) model can be interpreted as a supervised factor model, where two factor modelings are simultaneously applied to response and predictor spaces. This talk introduces a new model, called vector autoregression with common response and predictor factors, to explore further the common structure between the response and predictors in the VAR framework. The new model can provide better physical interpretations and improve estimation efficiency. In conjunction with the tensor operation, the model can easily be extended to any finite-order VAR model. A regularization-based method is considered for the high-dimensional estimation with the gradient descent algorithm, and its computational and statistical convergence guarantees are established. For data with pervasive cross-sectional dependence, a transformation for responses is developed to alleviate the diverging eigenvalue effect. Moreover, we consider additional sparsity structure in factor loading for the case of ultra-high dimension. Simulation experiments confirm our theoretical findings and a macroeconomic application showcases the appealing properties of the proposed model in structural analysis and forecasting.

Robust and optimal low-rank tensor estimation via robust gradient descent

Xiaoyu Zhang, Tongji University

Low-rank tensor models are prevalent in statistics and machine learning, but existing methods heavily rely on the sub-Gaussian distribution assumptions on data. To handle the heavy-tailed distributions in real applications, we propose a novel robust estimation procedure based on truncated gradient descent algorithm for the general low-rank tensor models. Computational convergence for the proposed method is provided with optimal statistical rates for various linear and nonlinear models. Specifically, the statistical errors are associated with the local moment condition, which characterizes the distributional properties of tensor variables projected onto certain low-dimensional local regions. Numerical studies are provided to show the efficacy of the proposed method.

Identifying Unobserved Functional Relationship in Traditional Data

Songhua Tan, Shanghai University of Finance and Economics

This paper proposes a novel functional linear quantile model for traditional time series data with an unobserved functional predictor. Specifically, the target conditional quantile function is affected by the whole conditional quantile function of the predictor. This proposed model can capture the functional relationship within two conditional distributions. We propose a simple and effective three-step approach of quantile regression estimation for our model based on functional principal component analysis. To the best of our knowledge, it is the first time that the functional principal component analysis is employed for traditional data, rather than functional data with full or partial observations, or panel data with cross-sectional observations. Unlike functional data analysis, which usually focuses on the convergence rate of the estimator, we investigate the asymptotic normality of the three-step estimator. A Monte Carlo experiment is conducted to investigate the finite sample performance of the proposed estimator and its asymptotic normality. The empirical study explores the upside and downside risk of GDP with economic condition, thereby offering valuable insights for policymakers.

Session 10	**Recent Advance in Dynamic Panel Models and Forecasting** *Organizer/Chair: Weining Wang, University of Groningen*	E22-2009	12 Dec, 11:00 - 12:30

Network regression with low rank and sparse structure

Yingxing Li, Xiamen University

We propose to study the interaction effects of social and spatial networks in the presence of a noisy adjacency matrix. First, we provide evidence that existing network datasets exhibit low-rank, sparse, and noisy structures, and we utilize this information to create a de-noised version of the network. We employ the Least Absolute Shrinkage and Selection Operator (LASSO) in conjunction with nuclear norm penalization to simultaneously regularize the sparse and low-rank components. We introduce two procedures: a two-step estimator and a supervised Generalized Method of Moments (GMM) estimator. We examine the large sample properties of our estimators. Simulation and application analysis indicate that our estimation methods perform favorably compared to standard GMM.

Uniform Inference on High-dimensional Spatial Panel Networks

Chen Huang, Aarhus University

We propose employing a debiased-regularized, high-dimensional generalized method of moments (GMM) framework to perform inference on large-scale spatial panel networks. In particular, network structure with a flexible sparse deviation, which can be regarded either as latent or as misspecified from a predetermined adjacency matrix, is estimated using debiased machine learning approach. The theoretical analysis establishes the consistency and asymptotic normality of our proposed estimator, taking into account general temporal and spatial dependency inherent in the data-generating processes. The dimensionality allowance in presence of dependency is discussed. A primary contribution of our study is the development of uniform inference theory that enables hypothesis testing on the parameters of interest, including zero or non-zero elements in the network structure. Additionally, the asymptotic properties for the estimator are derived for both linear and nonlinear moments. Simulations demonstrate superior performance of our proposed approach. Lastly, we apply our methodology to investigate the spatial network effect of stock returns.

Are asset pricing models sparse?

Jingyu He, City University of Hong Kong

We reveal the truth behind the illusion of sparsity by incorporating a conditional model within the Bayesian framework proposed by Geweke and Zhou (1996). Estimating factor models with a sparse representation in the high-dimensional space of factors and characteristics is a fundamental challenge in asset pricing. Our paper develops a novel Bayesian sparse latent conditional factor model that employs a spike-and-slab prior on characteristics to identify sparsity. Applying our method to the U.S. equity market, we provide robust evidence of sparsity and identify a definitive set of useful characteristics.

Estimation of Characteristics-based Quantile Factor Models

Liang Chen, Peking University

This paper studies the estimation of characteristics-based quantile factor models where the factor loadings are unknown functions of observed individual characteristics, while the idiosyncratic error terms are subject to conditional quantile restrictions. We propose a three-stage estimation procedure that is easily implementable in practice and has nice properties. The convergence rates, the limiting distributions of the estimated factors and loading functions, plus a consistent selection criterion for the number of factors at each quantile are derived under general conditions. The proposed estimation

methodology is shown to work satisfactorily when: (i) the idiosyncratic errors have heavy tails, (ii) the time dimension of the panel dataset is not large, and (iii) the number of factors exceeds the number of characteristics. Finite sample simulations and an empirical application aimed at estimating the loading functions of the daily returns of a large panel of S&P500 index securities help illustrate these properties.

Session 15	**Machine Learning Methods in Econometrics** *Organizer/Chair: Degui Li, University of Macau*	E22-2010	12 Dec, 11:00 - 12:30

Deep Conditional Distribution Learning via Conditional F\"{o}llmer Flow

Jinyuan Chang, Southwestern University of Finance and Economics

We introduce an ordinary differential equation (ODE) based deep generative method for learning conditional distributions, named Conditional F\"{o}llmer Flow. Starting from a standard Gaussian distribution, the proposed flow could approximate the target conditional distribution very well when the time is close to 1. For effective implementation, we discretize the flow with Euler's method where we estimate the velocity field nonparametrically using a deep neural network. Furthermore, we also establish the convergence result for the Wasserstein-2 distance between the distribution of the learned samples and the target conditional distribution, providing the first comprehensive end-to-end error analysis for conditional distribution learning via ODE flow. Our numerical experiments showcase its effectiveness across a range of scenarios, from standard nonparametric conditional density estimation problems to more intricate challenges involving image data, illustrating its superiority over various existing conditional density estimation methods.

Estimating Time-Varying Networks for High-Dimensional Time Series

Jia Chen, University of Macau

We explore time-varying networks for high-dimensional locally stationary time series, using the large VAR model framework with both the transition and (error) precision matrices evolving smoothly over time. Two types of time-varying graphs are investigated: one containing directed edges of Granger causality linkages, and the other containing undirected edges of partial correlation linkages. Under the sparse structural assumption, we propose a penalised local linear method with time-varying weighted group LASSO to jointly estimate the transition matrices and identify their significant entries, and a time-varying CLIME method to estimate the precision matrices. The estimated transition and precision matrices are then used to determine the time-varying network structures. Under some mild conditions, we derive the theoretical properties of the proposed estimates including the consistency and oracle properties. In addition, we extend the methodology and theory to cover highly-correlated large-scale time series, for which the sparsity assumption becomes invalid and we allow for common factors before estimating the factor-adjusted time-varying networks. We provide extensive simulation studies and an empirical application to a large U.S. macroeconomic dataset to illustrate the finite-sample performance of our methods.

Estimation of Large Dynamic Precision Matrices with a Latent Semiparametric Structure

Yuning Li, University of York

This paper studies the estimation of dynamic precision matrices with multiple conditioning variables for high-dimensional time series. We assume that the high-dimensional time series has an approximate factor structure plus an idiosyncratic error term, allowing the time series to have a non-sparse dynamic precision matrix and hence, enhancing the applicability of our method. Using the Sherman-Morrison-Woodbury formula, the estimation of the dynamic precision matrix for the time series boils down to the estimation of a low-rank factor structure and the precision matrix of the idiosyncratic error term. For the latter, we introduce an easy-to-implement semiparametric method to estimate the entries of the corresponding dynamic covariance matrix via the Model Averaging MArginal Regression (MAMAR) before applying the constrained L1 minimisation for inverse matrix estimation (CLIME) method to obtain the dynamic precision matrix. Under some regularity conditions, we derive the uniform consistency for the proposed estimators. We provide a simulation study that illustrates the finite-sample performance

of the developed methodology and an application in construction of minimum variance portfolios using daily returns of S&P 500 constituents from 2000 to 2023.

Optimal Treatment Allocation Strategies for A/B Testing in Two-sided Marketplaces

Chengchun Shi, London School of Economics

Time series experiments, in which experimental units receive a sequence of treatments over time, are frequently employed in many technological companies to evaluate the performance of a newly developed policy, product, or treatment relative to a baseline control. Many existing A/B testing solutions assume a fully observable experimental environment that satisfies the Markov condition, which often does not hold in practice. This paper studies the optimal design for A/B testing in partially observable environments. We introduce a controlled (vector) autoregressive moving average model to capture partial observability. We introduce a small signal asymptotic framework to simplify the analysis of asymptotic mean squared errors of average treatment effect estimators under various designs. We develop two algorithms to estimate the optimal design: one utilizing constrained optimization and the other employing reinforcement learning. We demonstrate the superior performance of our designs using a dispatch simulator and two real datasets from a ride-sharing company. A Python implementation of our proposal is available at https://github.com/datake/ARMADesign.

Session 16	**Recent Advances in Statistical Intelligence for Complex Data Analysis** *Organizer/Chair: Yuan Ke, University of Georgia*	E22-2011	12 Dec, 11:00 - 12:30

Robust estimation of number of factors in high dimensional factor modeling via Spearman's rank correlation matrix

Zeng Li, Southern University of Science and Technology

Determining the number of factors in high-dimensional factor modeling is essential but challenging, especially when the data are heavy-tailed. In this paper, we introduce a new estimator based on the spectral properties of Spearman's rank correlation matrix under the high-dimensional setting, where both dimension and sample size tend to infinity proportionally. Our estimator is applicable for scenarios where either the common factors or idiosyncratic errors follow heavy-tailed distributions. We prove that the proposed estimator is consistent under mild conditions. Numerical experiments also demonstrate the superiority of our estimator compared to existing methods, especially for the heavy-tailed case.

Some new perspectives for L_0 learning

Yuan Ke, University of Georgia

Many statistical learning problems involve identifying the optimal model from a subset of certain elements, and hence categorizing them under the domain of L_0 learning problems. Although solving L_0 learning problems is statistically attractive, it is also computationally expensive. In this presentation, we explore the potential of quantum computing as a revolutionary tool to tackle challenges associated with L_0 learning problems. We introduce a novel quantum optimization framework applicable to many L_0 learning applications. Our proposed algorithm not only offers a super-polynomial speed advantage over traditional methods used in electronic computing but is also demonstrated to be theoretically near-optimal. Comprehensive numerical experiments are conducted to demonstrate the finite sample performance of our quantum optimization technique. Additionally, we compare our method with popular classical algorithms, showcasing its efficacy across various L_0 learning applications.

Covariate-Adjusted Generalized Factor Analysis with Application to Testing Fairness

Ouyang Jing, The University of Hong Kong

Latent variable models are popularly used to measure latent factors (e.g., abilities and personalities) from large-scale assessment data. Beyond understanding these latent factors, the covariate effect on responses controlling for latent factors is also of great scientific interest and has wide applications, such as evaluating the fairness of educational testing, where the covariate effect reflects whether a test question is biased toward certain individual characteristics (e.g., gender and race), taking into account their latent abilities. However, the large sample sizes and test lengths pose challenges to developing efficient methods and drawing valid inferences. Moreover, to accommodate the commonly encountered discrete responses, nonlinear latent factor models are often assumed, adding further complexity. To address these challenges, we consider a covariate-adjusted generalized factor model and develop novel and interpretable conditions to address the identifiability issue. Based on the identifiability conditions, we propose a joint maximum likelihood estimation method and establish estimation consistency and asymptotic normality results for the covariate effects. Furthermore, we derive estimation and inference results for latent factors and the factor loadings. We illustrate the finite sample performance of the proposed method through extensive numerical studies and an educational assessment dataset from the Programme for International Student Assessment (PISA).

Session 17	**D4BI: Data-Driven Dynamic Decisions for Business Intelligence** *Organizer/Chair: Elynn Chen, New York University*	E22-2013	12 Dec, 11:00 - 12:30

Stock co-jump network with mixed memberships

Yingying Li, The Hong Kong University of Science and Technology

We propose a Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP) to study stock co-jump dependency. To estimate the community structure, we extend the SCORE algorithm in Jin (2015) and develop a Spectral Cluster- ing On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP) algorithm. We prove that SCORE-DMP enjoys strong consistency in community detection. Empirically, using high-frequency data of S&P 500 constituents, we construct two co-jump networks according to whether the market jumps and find that they exhibit different community features than GICS. We further show that the co-jump networks help in stock return prediction.

Inference for Changing Periodicity, Smooth Trend and Covariate Effects in Time Series

Lucy Xia, The Hong Kong University of Science and Technology

Traditional analysis of a periodic time series assumes its pattern remains the same. However, some recent empirical studies in climatology and other fields find that the amplitude may change over time, which has important implications. We develop a formal procedure to detect and estimate change-points in the periodic pattern. Often, there is also a smooth trend, and sometimes the period is unknown, with potential other covariate effects. Based on a new model that takes all of these factors into account, we propose a three-step estimation procedure to estimate them all accurately. First, we adopt penalized segmented least squares estimation for the unknown period, with the trend and covariate effects approximated by B-splines. Then, given the period estimate, we construct a novel SupF statistic and use it in binary segmentation to estimate change-points in the periodic component. Finally, given the period and change-point estimates, we estimate the entire periodic component, trend, and covariate effects. Asymptotic results for the proposed estimators are derived, including consistency of the period and change-point estimators, and the asymptotic normality of the estimated periodic sequence, trend and covariate effects. Simulation results demonstrate the appealing performance of the new method, while empirical studies highlight its advantages.

Balancing Utility and Cost in Dynamic Treatment Regimes

Yuqian Zhang, Renmin University of China

Dynamic treatment regimes (DTRs) refer to personalized, adaptive treatment strategies designed to guide the sequential allocation of treatments for individuals over time. At each stage, individual characteristics are collected to facilitate more precise adjustments to treatment plans. However, in practice, collecting features that accurately reflect an individual's state often incurs higher costs. In this work, we propose a new strategy for estimating DTRs that accounts for the costs associated with feature collection over time. Our method is doubly robust and accommodates high-dimensional covariates and non-parametric nuisance estimates. The performance of the proposed method is demonstrated through extensive numerical studies.

Context-Based Dynamic Pricing with Separable Demand Models (joint work with David Simchi-Levi, and Chonghuan Wang)

Jinzhi Bu, The Hong Kong Polytechnic University

Motivated by the empirical evidence observed from the real-world dataset, we consider a context-based dynamic pricing problem with separable demand models. The demand function is endowed with a

separable structure in the form of f(p)+g(x), where p and x denote the price and feature vector respectively. Under different configurations of f(p) and g(x), we systematically characterize the statistical complexity of the online learning problem. Specifically, we consider three models: (i) f(p) is linear and g(x) is non-parametric; (ii) f(p) is non-parametric and g(x) is linear; and (iii) f(p) and g(x) are both non-parametric. For each model, we design an efficient learning algorithm with a provable regret upper bound and establish an almost matching regret lower bound.

Dynamic Contextual Pricing with Doubly Non-Parametric Random Utility Models

Jiayu Li, New York University

In the evolving landscape of digital commerce, adaptive dynamic pricing strategies are essential for gaining a competitive edge. This paper introduces novel {\em doubly nonparametric random utility models} that eschew traditional parametric assumptions used in estimating consumer demand's mean utility function and noise distribution. Existing nonparametric methods like multi-scale {\em Distributional Nearest Neighbors (DNN and TDNN)}, initially designed for offline regression, face challenges in dynamic online pricing due to design limitations, such as the indirect observability of utility-related variables and the absence of uniform convergence guarantees. We address these challenges with innovative population equations that facilitate nonparametric estimation within decision-making frameworks and establish new analytical results on the uniform convergence rates of DNN and TDNN, enhancing their applicability in dynamic environments. Our theoretical analysis confirms that the statistical learning rates for the mean utility function and noise distribution are minimax optimal. We also derive a regret bound that illustrates the critical interaction between model dimensionality and noise distribution smoothness, deepening our understanding of dynamic pricing under varied market conditions. These contributions offer substantial theoretical insights and practical tools for implementing effective, data-driven pricing strategies, advancing the theoretical framework of pricing models and providing robust methodologies for navigating the complexities of modern markets.

Session 19	**Flexible Data Science Learning Inference** *Organizer/Chair:* *Jinchi Lv, University of Southern California;* *Yingying Fan, University of Southern California*	E22-2014	12 Dec, 11:00 - 12:30

Berry-Esseen Bounds For Degenerate U-Statistic With Application To the Distance Correlation

Qiman Shao, Southern University of Science and Technology

Let $X_1, X_2, . . . ,X_n$ be independent and identically distributed random vectors, $ T_n = T_n(X_1,X_2, . . . ,X_n)$ be a degenerate U-statistic, and $\Delta_n = Δ_n(X_1,X_2, . . . ,X_n)$ be a remainder term. In this paper, we establish a Berry-Esseen type theorem for $T_n +\Delta_n$ by an exchangeable pair approach. As an application, a sharp error bound of normal distribution approximation for the distance correlation is obtained, which improves some results in Gao, Fan, Lv and Shao (2021). This talk is based on a joint work with Songhao Liu and Hao Shi.

Online Nonparametric Learning

Fang Yao, Peking University

Online learning and modeling has attracted considerable interest due to increasingly available data in streaming manner. Nonparametric models, although flexible, have seen limited use in online settings due to their data-driven nature and high computational demands. We introduce an innovative online method for dynamically updating local polynomial regression estimates. Our approach decomposes kernel-type estimates into two sufficient statistics and approximates future optimal bandwidths with a dynamic candidate sequence. This idea extends to general nonlinear optimization problems, where we propose an online smoothing backfitting algorithm for generalized additive models (GAM). We establish asymptotic normality and efficiency lower bounds for online estimation, shedding light on the trade-off between accuracy and computational cost driven by the bandwidth sequence length. For GAM, We also investigate statistical and algorithmic convergence and provide a framework for balancing estimation and computation performance. Our proposed online estimation is also applicable to complex structural data such as functional data. Simulations and real data examples are provided to support the usefulness of the proposed method.

Robust Knockoffs Inference with Coupling

Yingying Fan, University of Southern California

We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and k-familywise error rate (k-FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or k-FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework. Additionally, we formally connect our concept of knockoff variable coupling to a type of Wasserstein distance.

SOFARI: High-Dimensional Manifold-Based Inference

Zemin Zheng, University of Science and Technology of China

Multi-task learning is a widely used technique for harnessing information from various tasks. Recently, the sparse orthogonal factor regression (SOFAR) framework, based on the sparse singular value decomposition (SVD) within the coefficient matrix, was introduced for interpretable multi-task learning, enabling the discovery of meaningful latent feature-response association networks across different layers. However, conducting precise inference on the latent factor matrices has remained challenging due to the orthogonality constraints inherited from the sparse SVD constraints. In this paper, we suggest a novel approach called the high-dimensional manifold-based SOFAR inference (SOFARI), drawing on the Neyman near-orthogonality inference while incorporating the Stiefel manifold structure imposed by the SVD constraints. By leveraging the underlying Stiefel manifold structure that is crucial to enabling inference, SOFARI provides easy-to-use bias-corrected estimators for both latent left factor vectors and singular values, for which we show to enjoy the asymptotic mean-zero normal distributions with estimable variances. We illustrate the effectiveness of SOFARI and justify our theoretical results through simulation examples and a real data application in economic forecasting.

Session 20	**Statistics of Machine Learning** *Organizer/Chair:* *Yingcun Xia, National University of Singapore;* *Qian Lin, Tsinghua University*	E22-2015	12 Dec, 11:00 - 12:30

Frechet Cumulative Covariance Net for Nonlinear Sufficient Dimension Reduction

Zhou Yu, East China Normal University

We introduce the Frechet Cumlative Covariace as a dependence measure between metric space valued response and Euclidean predictors. And we further demonstrate that the Frechet Cumuative Covariance can be adopted to identify the central sigmal field for nonlinear sufficient dimenison reduction. In the sample level, the Frechet Cumulative Covariance net based on MLP or CNN is developed. And we present the nonasymptotic error bound for our proposed estimator. Comprehensive numerical studies demonstrate the efficiency of our proposal.

On non-redundant and linear operator-based nonlinear dimension reduction

Wei Luo, Zhejiang University

Kernel principal component analysis (KPCA), a popular nonlinear dimension reduction technique, aims at finding a basis of a presumed low-dimensional function space. This causes the redundancy issue that each kernel principal component can be a measurable function of the preceding components, which harms the effectiveness of dimension reduction and leaves the dimension of the reduced data a heuristic choice. In this paper, we formulate the parameter of interest for nonlinear dimension reduction as a small function set that generates the σ-field of the original data, and, using a novel characterization of near conditional mean independence, we propose two sequential dimension reduction methods that address the redundancy issue, have the same level of computational complexity as KPCA, and require more plausible assumptions on the singularity of the original data. Compared with the other nonlinear dimension reduction methods, the proposed methods are applicable to various complex cases with guarantee on both the asymptotic consistency and the smoothness and interpretability of the reduced data. By constructing a measure of exhaustiveness of the reduced data, we also provide consistent order determination for these methods. Some supportive numerical studies are presented at the end.

Toward a Unified Non-Euclidean Semiparametric Efficiency Theory

Lyufang Sun, National University of Singapore

Responding the pressing challenge of analyzing increasingly complex datasets, particularly those with non-Euclidean/nonlinear structures, many novel statistical models and associated inferential procedures have been proposed in recent years. However, unlike the counterpart when the underlying space is linear, the corresponding efficiency theory has not been comprehensively developed.
For concreteness, we ground our work by considering a particular nonlinear space -- the Riemmanian manifold. In particular, We generalize the notions of regular estimator, local asymptotic normality and differentiable functionals. Further, we develop the generalized H\'{a}jek-Le Cam convolution theorem to be the theoretical justification of semiparametric efficient estimation in nonlinear spaces. We also demonstrate the wide applicability of our framework by applying it to examples including single index model, image registration and geodesically convex M-estimation

Mathematical Theory for Deep Learning

Qian Lin, Tsinghua University

We will provide a brief review of what we have done in kernel regression (a.k.a., the NTK theory) and propose a novel framework to explain the superiority of deep neural network.

Session 32	**Financial Big Data** *Organizer/Chair: Yingying Li, The Hong Kong University of Science and Technology*	E22-2017	12 Dec, 11:00 - 12:30

Sub-Gaussian Estimation of the Scatter Matrix under High-Dimensional Elliptical Factor Model with 2+epsilonth Moment

Xinghua Zheng, The Hong Kong University of Science and Technology

We study the estimation of high-dimensional scatter matrices under elliptical factor models with 2+epsilonth moment. For such heavy-tailed data, robust estimators like the Huber-type estimator in Fan, Liu, and Wang (2018) can not achieve a sub-Gaussian convergence rate. In this paper, we develop an idiosyncratic-projected self-normalization method to remove the effect of the heavy-tailed scalar component and propose a robust estimator of the scatter matrix that achieves the sub-Gaussian rate. The estimator demonstrates superior performance in estimating high-dimensional global minimum variance portfolios. Based on joint work with Yi Ding.

Navigating Complexity: Constrained Portfolio Analysis in High Dimensions with Tracking Error and Weight Constraints

Qingliang Fan, The Chinese University of Hong Kong

This paper explores the statistical properties of forming constrained optimal portfolios within a high-dimensional set of assets. We examine portfolios with tracking error constraints, those with simultaneous tracking error and weight restrictions, and portfolios constrained solely by weight. Tracking error measures portfolio performance against a benchmark (typically an index), while weight constraints determine asset allocation based on regulatory requirements or fund prospectuses. Our approach employs a novel statistical learning technique that integrates factor models with nodewise regression, named the Constrained Residual Nodewise Optimal Weight Regression (CROWN) method. We demonstrate its estimation consistency in large dimensions, even when assets outnumber the portfolio's time span. Convergence rate results for constrained portfolio weights, risk, and Sharpe Ratio are provided, and simulation and empirical evidence highlight the method's outstanding performance.

Does Noise Hurt Economic Forecasts?

Zhentao Shi, The Chinese University of Hong Kong

This paper explores whether variable selection enhances economic forecasting. While economists often remove noise from predictors, we show that economic forecast models are not sparse if the outcome is driven by latent factors. We also prove a compelling result that including noise in predictions yields greater benefits than excluding it. Empirically, we apply this method to four common forecasting applications including forecasting the U.S. inflation rate and obtain results that surpass many commonly used models that rely on dimension reduction or variable selection.

Session 45	**Novel Machine Learning and Network Models** *Organizer/Chair:* *Xinyuan Song, The Chinese University of Hong Kong;* *Xinzhou Guo, The Hong Kong University of Science and Technology*	E22-2018	12 Dec, 11:00 - 12:30

Word-Level Maximum Mean Discrepancy Regularization for Word Embedding

Ben Dai, The Chinese University of Hong Kong

The technique of word embedding is widely used in natural language processing (NLP) to represent words as numerical vectors in textual datasets. However, the estimation of word embedding may suffer from severe overfitting due to the huge variety of words. To address the issue, this article proposes a novel regularization framework that recognizes and accounts for the "word-level distribution discrepancy"–a common phenomenon in a range of NLP tasks where word distributions are noticeably disparate under different labels. The proposed regularization, referred to as word-level MMD (wMMD), is a variant of maximum mean discrepancy (MMD) that serves a specific purpose: to enhance/preserve the distribution discrepancies within word embedding numerical vectors and thus prevent overfitting. Our theoretical analysis illustrates that w MMD can effectively operate as a dimension reduction technique of word embedding, thereby significantly improving the robustness and generalization of NLP models. The numerical effectiveness of w MMD is demonstrated in various simulated examples, CE-T1 and BBC News datasets with state-of-the-art NLP deep learning architectures.

Federated Double Machine Learning for High-dimensional Semiparametric Regression

Kai Kang, Sun Yat-sen University

Federated learning trains a global model while keeping data local, but current methods struggle with high-dimensional semiparametric models involving complex nuisance parameters. The aim of this paper is to propose a Federated Double Machine Learning (FDML) framework for high-dimensional semiparametric problems in multicenter studies. Our approach leverages double machine learning (Chernozhukov et al., 2018a) to estimate center-specific parameters, extends the surrogate efficient score method to a Neyman-orthogonal setting, and applies density ratio tilting to create a federated estimator that uses local individual-level data and summary statistics from other centers. This mitigates regularization bias and overfitting in high-dimensional nuisance estimation. We establish the estimator's limiting distribution under minimal assumptions, validate its accuracy through extensive simulations, and demonstrate its effectiveness on the analysis of multiple phase data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

Symmetric Graph Convolutional Auto-encoder for Scalable and Accurate Study of Spatial Transcriptomics

Zhixiang Lin, The Chinese University of Hong Kong

Recent advances in spatial transcriptomics (ST) have enabled comprehensive profiling of gene expression with spatial information in the context of the tissue microenvironment. However, with the improvements in the resolution and scale of ST data, deciphering spatial domains precisely while ensuring efficiency and scalability is still challenging. Here, we develop SGCAST, an efficient auto-encoder framework to identify spatial domains. SGCAST adopts a symmetric graph convolutional auto-encoder to learn aggregated latent embeddings via integrating the gene expression similarity and the proximity of the spatial spots. This framework in SGCAST enables a mini-batch training strategy, which makes SGCAST memory-efficient and scalable to high-resolution spatial transcriptomic data with a large number of spots. SGCAST improves the overall accuracy of spatial domain identification on benchmarking data. We also validated the performance of SGCAST on ST datasets at various scales

across multiple platforms. Our study illustrates the superior capacity of SGCAST on analyzing spatial transcriptomic data.

Survival Mixed Membership Blockmodel for Time-to-event Data on Social Networks

Fangda Song, The Chinese University of Hong Kong, Shenzhen

Whenever we send a message via a channel such as E-mail, Facebook, WhatsApp, WeChat, or LinkedIn, we care about the response rate—the probability that our message will receive a response—and the response time—how long it will take to receive a reply. Recent studies have made considerable efforts to model the sending behaviors of messages in social networks with point processes. However, statistical research on modeling response rates and response times on social networks is still lacking. Compared with sending behaviors, which are often determined by the sender's characteristics, response rates and response times further depend on the relationship between the sender and the receiver. Here, we develop a survival mixed membership blockmodel (SMMB) that integrates semiparametric cure rate models with a mixed membership stochastic blockmodel to analyze time-to-event data observed for node pairs in a social network, and we are able to prove its model identifiability without the pure node assumption. We develop a Markov chain Monte Carlo algorithm to conduct posterior inference and select the number of social clusters in the network according to the conditional deviance information criterion. The application of the SMMB to the Enron E-mail corpus offers novel insights into the company's organization and power relations.

Session 53	**Corporate Information and Market Pricing Mechanism** *Organizer/Chair: Jing Xie, University of Macau*	E22-G008	12 Dec, 11:00 - 12:30

Corporate Finance Through Loyalty Programs

Dan Luo, The Chinese University of Hong Kong

Loyalty programs (LPs) are widely prevalent and typically analyzed in economic research for their role in boosting income. This paper uncovers a novel role of LPs as financing instruments. The rewards issued to and redeemed by consumers cause shifts in firms' present and future cash flows, effectively creating a form of borrowing from consumers. We document three stylized facts about LPs in the airline and hotel industries: 1) LPs serve as significant financing sources, with co-branded credit card programs contributing a large portion; 2) rewards are issued through broad consumption but are redeemed predominantly for consumption related to the issuing firm; 3) LPs generate countercyclical cash flows. We then build a dynamic model of LPs as financing instruments. The model features convenient rewards, which consumers can freely redeem. As a result, the funds raised through LPs emerge endogenously in equilibrium as a result of the interplay between reward issuance and redemption. The model suggests that 1) firms supplying high-value, low-frequency services can leverage LPs more effectively for financing; 2) LP financing has special cyclical natures that are attractive to highly cyclical firms; 3) firms should aim to decouple reward issuance from their business; 4) firms should limit consumers' discretion to purchase or transfer rewards.

A General Test for Functional Inequalities

Wenyu Zhou, Zhejiang University

This paper develops a nonparametric test for general functional inequalities that include conditional moment inequalities as a special case. It is shown that the test controls size uniformly over a large class of distributions for observed data, importantly allowing for general forms of time series dependence. New results on uniform growing dimensional Gaussian coupling for general mixingale processes are developed for this purpose, which readily accommodate most applications in economics and finance. The proposed method is applied in a portfolio evaluation context to test for "all-weather" portfolios with uniformly superior conditional Sharpe ratio functions.

STEM Auditors and Financial Reporting Timeliness

Shaohua Tian, Macao Polytechnic University

Timeliness is an enhancing qualitative characteristic that augments the relevance of information presented in financial reporting. Auditors are pivotal in the preparation and issuance of financial statements. Despite existing research examining the impact of auditors' educational background on audit efficiency, the results are inconsistent. This research aims to investigate the influence of auditors with an educational background in Science, Technology, Engineering, and Mathematics (STEM) on the timeliness of financial reporting. Using a sample of Chinese listed companies from 2001 to 2022, we find that companies with STEM-educated lead auditors tend to report their financial statements in a timely fashion. Furthermore, the increase in timeliness does not compromise reporting quality. The results imply a potential advantage conferred by STEM education in the audit profession, which may necessitate a reevaluation and enhancement of accounting curricula within higher education institutions.

The Liquidity Premium and Long-Run Risk Before and After 2000

Nan Li, Shanghai Jiao Tong University

This paper examines whether changes in the liquidity premium before and after 2000 can be explained by shifts in exposures to long-run risk. We estimate the risk exposures of liquidity-based portfolios to

consumption and investment long-run risk factors, as identified in Li's (2024) long-run risk model, which builds on the model of Hansen, Heaton, and Li (2008) by incorporating both aggregate and investment-specific technological shocks. We find that illiquid stocks exhibit significantly larger exposures to both risk factors than liquid stocks over the full sample period from 1964 to 2023 and in the pre-2000 subsample, consistent with the positive liquidity premium observed during these intervals. After 2000, the negative exposure of illiquid stocks to the long-run consumption risk factor accounts for the diminishing liquidity premium observed in this period.

Session 47	**Recent Advances in Complex Data Analysis** *Organizer/Chair: Yang Feng, New York University* *Huihang Liu, University of Science and Technology of China*	E22-G004	12 Dec, 11:00 - 12:30

Optimal Sparse Sliced Inverse Regression via Random Projection

Xin Chen, Southern University of Science and Technology

We propose a novel sparse sliced inverse regression method based on random projections in a large p small n setting. Embedded in a generalized eigenvalue framework, the proposed approach finally reduces to parallel execution of low-dimensional (generalized) eigenvalue decompositions, which facilitates high computational efficiency. Theoretically, we prove that this method achieves the minimax optimal rate of convergence under suitable assumptions. Furthermore, our algorithm involves a delicate reweighting scheme, which can significantly enhance the identifiability of the active set of covariates. Extensive numerical experiments demonstrate high superiority of the proposed algorithm in comparison to competing methods.

Post selection inference for censored quantile regression

Tony Sit, The Chinese University of Hong Kong

This paper proposes a novel method for constructing confidence intervals for censored quantile regression in high-dimensional data settings where the number of covariates may significantly exceed the sample size. Building on the weighted loss function introduced by Wang and Wang (2009; Sinica), we apply an L1 penalisation and subsequently perform a debiasing process on the resulting estimate. The debiased estimator is shown to exhibit asymptotic normality, providing a robust basis for inference. Unlike existing research, our approach relaxes the global linearity condition to a local linearity condition near the quantile of interest, offering a more flexible and accurate model. This method is particularly advantageous when dealing with heteroskedastic effects or violations of global linearity. Simulation results demonstrate superior performance of our method in constructing confidence intervals.

Alocal perspective in latent space network models

Lijia Wang, City University of Hong Kong

The influence of neighborhood effects in social networks is profound, affecting individual decision-making, opinion formation, and other personal dynamics. Thus, to understand the role of social networks in shaping individual behaviors and attitudes, it is important to begin with an understanding of an individual's localized viewpoint within the global network context. In this paper, we consider a general latent space network model and the problem of inferring the latent positions of the nodes, utilizing only a partial information network centered at a given individual. We use the novel individual-centered partial information framework to characterize the individual local view and a projected gradient descent algorithm for the parameter estimation. We further demonstrate that the convergence rate of our estimates is influenced by the characteristics of the node's neighborhood structure. We accordingly introduce a metric to quantify the bias present in the individual's local view.

Session 18	**New Statistical Methods and Modeling for Data Heterogeneity, Security, and Surveys** *Organizer/Chair:* *Chunming Zhang, University of Wisconsin Madison;* *Lei Wang, Nankai University*	E22-G015	12 Dec, 11:00 - 12:30

A GMM approach in coupling internal data and external summary information with heterogeneous data populations

Lei Wang, Nankai University

Because of advances in data collection and storage, statistical analysis in modern scientific research and practice now has opportunities to utilize external information such as summary statistics from similar studies. A likelihood approach based on a parametric model assumption has been developed in the literature to utilize external summary information when the populations for external and main internal data are assumed to be the same. In this article, we instead consider the generalized estimation equation (GEE) approach for statistical inference, which is semiparametric or nonparametric, and show how to utilize external summary information even when internal and external data populations are not the same. Our approach is coupling the internal data and external summary information to form additional estimation equations and then applying the generalized method of moments (GMM). We show that the proposed GMM estimator is asymptotically normal and, under some conditions, is more efficient than the GEE estimator without using external summary information. Estimators of the asymptotic covariance matrix of the GMM estimators are also proposed. Simulation results are obtained to confirm our theory and quantify the improvements by utilizing external data. An example is also included for illustration.

Modeling and Predicting Data Breach Risks

Peng Zhao, Jiangsu Normal University

In recent years, data breaches have become a significant concern, leading to substantial financial losses annually. However, the lack of suitable statistical approaches for assessing breach risks poses an obstacle. To address this challenge, we first propose a novel statistical model that focuses on analyzing hacking breach risks at the individual company level. We then develop a multivariate frequency-severity framework that examines breach risks at the state level. Additionally, we introduce the concept of the data breach lifecycle. By incorporating this lifecycle and utilizing a novel self-exciting marked point process model, we enhance our understanding of the temporal dynamics of data breaches. Applications in insurance industry are also presented.

Augmented two-step estimating equations with nuisance functionals and complex survey data

Puying Zhao, Yunnan University

Statistical inference in the presence of nuisance functionals with complex survey data is an important topic in social and economic studies. The Gini index, Lorenz curves and quantile shares are among the commonly encountered examples. The nuisance functionals are usually handled by a plug-in nonparametric estimator and the main inferential procedure can be carried out through a two-step generalized empirical likelihood method. Unfortunately, the resulting inference is not efficient and the nonparametric version of the Wilks' theorem breaks down even under simple random sampling. We propose an augmented estimating equations method with nuisance functionals and complex surveys. The second-step augmented estimating functions automatically handle the impact of the first-step plug-in estimator, and the resulting estimator of the main parameters of interest is invariant to the first step method. More importantly, the generalized empirical likelihood based Wilks' theorem holds for the main

parameters of interest under the design-based framework for commonly used survey designs, and the maximum generalized empirical likelihood estimators achieve the semiparametric efficiency bound. Performances of the proposed methods are demonstrated through simulation studies and an application using the dataset from the New York City Social Indicators Survey.