Invited Sessions II – 11 December 2024, 16:30 – 18:00

Session 02	**Recent Developments in Modern Statistical Learning** *Organizer/Chair: Yingying Fan, University of Southern California*	E22-2002	11 Dec, 16:30 - 18:00

Transfer Q*-Learning: Stationary and Non-Stationary MDPs

Elynn Chen, New York University

In dynamic decision-making scenarios across business, healthcare, and education, leveraging data from diverse populations can significantly enhance reinforcement learning (RL) performance for specific target populations, especially when target samples are limited. We develop comprehensive frameworks for transfer learning in RL, addressing both stationary Markov decision processes (MDPs) with iterative Q-learning and non-stationary finite-horizon MDPs with backward inductive learning.

For stationary MDPs, we propose an iterative Q-learning algorithm with knowledge transfer, establishing theoretical justifications through faster convergence rates under similarity assumptions. For non-stationary finite-horizon MDPs, we introduce two key innovations: (1) a novel "re-weighted targeting procedure" that enables vertical information-cascading along multiple temporal steps, and (2) transferred deep Q^*-learning that leverages neural networks as function approximators. We demonstrate that while naive sample pooling strategies may succeed in regression settings, they fail in MDPs, necessitating our more sophisticated approach. We establish theoretical guarantees for both settings, revealing the relationship between statistical performance and MDP task discrepancy. Our analysis illuminates how source and target sample sizes impact transfer effectiveness. The framework accommodates both transferable and non-transferable transition density ratios while assuming reward function transferability. Our analytical techniques have broader implications, extending to supervised transfer learning with neural networks and domain shift scenarios. Empirical evidence from both synthetic and real datasets validates our theoretical results, demonstrating significant improvements over single-task learning rates and highlighting the practical value of strategically constructed transferable RL samples in both stationary and non-stationary contexts.*

Related papers:
Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning: Manuscript on demand.*
Data-Driven Knowledge Transfer in Batch Q Learning: https://arxiv.org/abs/2404.15209*
Transfer Q-Learning: https://arxiv.org/abs/2202.04709

A Multi-Scale Leverage Effect Estimator with Dependent Microstructure Noise

Christina Dan Wang, NYU-Shanghai

We introduce a novel estimator for the leverage effect in the presence of microstructure noise, characterized by long-term dependence and higher-order moments in ultra-high-frequency data. Within the general Itô semimartingale framework, we first develop estimators applicable across various subsampling frequencies and establish their asymptotic properties. Subsequently, we combine leverage effect estimators from multiple frequencies to enhance the convergence rate. We establish the asymptotic properties and feasible CLT for the new method, termed the Multi-Time Scale Leverage Effect Estimator (MSLE). For MSLE, the asymptotic variance contributions from noise and discretization are additive rather than multiplicative, as seen in pre-averaging methods. This additive property allows for a data-driven approach in selecting estimator parameters, thus achieving superior performance in finite samples. Monte Carlo simulations demonstrate the strong finite-sample performance of our estimators.

Asymptotics of Yule-Walker Estimators for ARH(p) Model

Zhao Chen, Fudan University

We extend the classical AR models to Hilbert space with the ARH(p) model, enhancing the capability to analyze functional data. Based on the introduced autocovariance operator, we establish Yule-Walker equations for estimating parameters and create predictors for future observations within this framework. After ensuring model identifiability, we address the inverse problem through regularization, leading to consistent estimators and predictors. Our theoretical contributions are concentrated on deriving the Central Limit Theorem (CLT) for predictors, which enables the construction of confidence intervals and facilitates statistical inference. The mean squared prediction error is analyzed, highlighting the trade-off between bias and variance. We validate our findings through numerical simulations and apply the model to wearable device data, demonstrating distinct autoregressive operator characteristics across leisure activities, which enable effective classification with SVM.

Session 04	**Recent Advances in Econometrics** *Organizer/Chair: Qingfeng Liu, Hosei University*	E22-2007	11 Dec, 16:30 - 18:00

Robust Reproducible Network Exploration

Yoshimasa Uematsu, Hitotsubashi University

We propose a novel method of network detection that is robust against any complex dependence structure. Our goal is to conduct exploratory network detection, meaning that we attempt to detect a network composed of ``connectable" edges that are worth investigating in detail for further modelling or precise network analysis. For a reproducible network detection, we pursue high power while controlling the false discovery rate (FDR). In particular, we formalize the problem as a multiple testing, and propose p-variables that are used in the Benjamini-Hochberg procedure. We show that the proposed method controls the FDR under arbitrary dependence structure with any sample size, and has an asymptotic power one. The validity is also confirmed by simulations and a real data example.

Detecting structural breaks in spatial panel data models with unknown networks

Ryo Okui, University of Tokyo

This paper aims to detect structural break points in latent networks in a panel data setting. We consider panel models where the outcome of a unit depends on the outcomes and characteristics of other units. The latent network structure induces high-dimensional parameters and interactive outcomes generate endogeneity. Our goal is to detect breaks in high-dimensional network parameters associated with endogenous variables. We propose a two-step penalized nonlinear least squares approach to estimate the break points based on reduced forms, and show that the resulting estimator achieves superconsistency. This property allows us to estimate, and make inferences on, network and slope parameters as if the true break points were known.

Functional PCA for Surface Time Series

Yasumasa Matsuda, Tohoku University

Surface time series is a kind of spatial panel data, i.e., panel data when a cross-sectional unit is spatially observed. In surface time series analysis, cross-sectional unit usually locates irregularly with lots of NAs, resulting in a so-called unbalanced panel data . It follows that usual PCA which assumes complete panels does not work because of the features. In this talk, we regard the surface time series as a time series of functional data, for which functional principal component analysis (fPCA) is introduced with cubic B-spline basis estimation. We provide fPCA for surface time series with theoretical justifications and demonstrate how it works empirically for real surface time series.

Tying Maximum Likelihood Estimation for Dependent Data

Qingfeng Liu, Hosei University

This study proposes a tying maximum likelihood estimation (TMLE) method to improve the estimation performance of statistical and econometric models in which most time series have long sample periods and the rest have noticeably short sample periods. Essentially, the TMLE ties together the parameters of the long time series with those of the short time series so that useful information from the long time series can be transferred to the short ones. This information can help improve the estimation accuracy of the parameters related to the short series. We present the asymptotic properties of the TMLE and show its finite-sample risk bound under a fixed tuning parameter that determines the strength of the tying. Further, we present a method for selecting the tuning parameter based on a bootstrapping procedure. Finite-sample theories on this selection method are derived to describe how to effectively

conduct the bootstrapping procedure. Extensive artificial simulations and empirical applications show that the TMLE exhibits outstanding performance in point estimates and forecasts.

Session 43	**Statistical Learning for Econometric Models** *Organizer/Chair: Wei Zhong, Xiamen University*	E22-2009	11 Dec, 16:30-18:00

Differentially Private Sliced Inverse Regression: Minimax Optimality and Algorithm

Zhanrui Cai, The University of Hong Kong

Privacy preservation has become a critical concern in high-dimensional data analysis due to the growing prevalence of data-driven applications. Since its proposal, sliced inverse regression has been one of the most popular technique for reducing covariate dimensionality while maintaining sufficient statistical information. In this paper, we propose optimally differentially private algorithms specifically designed to address privacy concerns in the context of sufficient dimension reduction. We establish lower bounds for differentially private sliced inverse regression in both the low and high-dimensional settings. Moreover, we develop differentially private algorithms that achieve the minimax lower bounds up to logarithmic factors. Through a combination of simulations and real data analysis, we illustrate the efficacy of these differentially private algorithms in safeguarding privacy while preserving vital information within the reduced dimension space. As a natural extension, we can readily offer analogous lower and upper bounds for differentially private sparse principal component analysis, a topic that may also be of potential interest to the statistical and machine learning community.

A consistent specification test for expectile models

Xiaojun Song, Peking University

In this article, we propose a nonparametric test for the correct specification of parametric expectile models over a continuum of expectile levels. The test is based on continuous functionals of a residual-marked empirical process. We show that the test is consistent and has nontrivial power against a sequence of local alternatives approaching the null at a parametric rate. Since the limiting distribution of the test statistic is nonpivotal, we propose a simple multiplier bootstrap procedure to approximate the critical values. A Monte Carlo study shows that the asymptotic results provide good approximations for small sample sizes.

Statistical ranking with dynamic covariates

Ruijian Han, The Hong Kong Polytechnic University

We consider a covariate-assisted ranking model within the Plackett–Luce framework. Unlike previous works focusing on pure covariates or individual effects with fixed covariates, our approach integrates individual effects with dynamic covariates. This increased flexibility enhances model fitting by allowing for individualized dynamic ranking but also presents significant challenges in theoretical analysis. This paper addresses these challenges in the context of maximum likelihood estimation (MLE). We begin by providing sufficient and necessary conditions for both model identifiability and the unique existence of the MLE. Then, we propose an alternating maximization algorithm to compute the MLE. Under suitable assumptions on the graph topology, we establish uniform consistency for the MLE with convergence rates characterized by the asymptotic graph connectivity. The proposed graph topology assumption holds in several random graph models with optimal leading-order sparsity. Comprehensive numerical studies are conducted to corroborate our findings and demonstrate the application of the proposed model to real-world datasets. This is a joint work with Pinjun Dong, Binyan Jiang and Yiming Xu.

Modelling Homophily in Dynamic Networks

Binyan Jiang, The Hong Kong Polytechnic University

Statistical modeling of network data is an important topic in various areas. Although many real networks are dynamic in nature, most existing statistical models and related inferences for network data are *confined to static networks, and the development of the foundation for dynamic network models is still in its infancy. In particular, to the best of our knowledge, no attempts have been made to jointly address node heterogeneity and link homophily among dynamic networks. Being able to capture these network features simultaneously will only bring new insights on understanding how networks were formed, but also provide more sophisticated tools for the prediction of a future network with statistical guarantees. In this paper, we take into account link homophily associated with both observed traits and latent traits of the nodes and propose a novel convex loss-based framework to generate stable estimations for the high dimensional parameters. We show that, with an appropriate initialization, the resulting estimator is consistent. The promising performance of our proposed model is demonstrated through its application in community detection as well as various simulation studies.*

Session 06	**Network Autoregression and Time Series Learning** *Organizer/Chair: Chao Zheng, University of Southampton*	E22-2010	11 Dec, 16:30 - 18:00

Estimation of Financial Network by Frequentist Model Averaging

Huihang Liu, University of Science and Technology of China

Advances in information technologies have made network data increasingly frequent in finance. To estimate the financial network, we propose an optimal model averaging method for directed acyclic Gaussian graphs. With a set of candidate models varying by graph structures, we average estimates from candidate models using weights that minimize a penalized negative log-likelihood criterion. We not only build the asymptotic optimality, weight convergence, and parameter consistency for the proposed method, but also, we clarify the impact of different models on the convergence rate and prove the parameter consistency under misspecified candidate graph models. Results of simulation studies and a real-data analysis on banks' international liability data show the promise of the proposed method.

Functional quantile auto regression

Chaohua Dong, Zhongnan University of Economics and Law

This paper proposes a new class of time series models, the functional quantile autoregression (FQAR) models, in which the conditional distribution of the observation at the current time point is affected by its past distributional information, and is expressed as a functional of the past conditional quantile functions. Different from the conventional functional time series models which are based on functionally observed data, the proposed FQAR method studies functional dynamics in traditional time series data. We propose a sieve estimator for the model. Asymptotic properties of the estimators are derived. Numerical investigations are conducted to highlight the proposed method.

Deep Neural Network Estimation for Non-IID Data

Chao Zheng, University of Southampton

Theoretical development of deep neural networks has been heavily investigated in recent years, and most works have been focused on the setting of independent data. In this work, we study the theoretical properties of deep feedforward ReLU networks on modelling non-linear non-IID mixing sequence, which includes a wide range of time series models such as AR process. We present non-asymptotic generalization bounds for the estimation error and show that it is related to the dependence structure in the data and the architect of DNN.

Statistical Random Forests for Nonlinear Time Series Modelling

Shihao Zhang, University of Southampton

Random forests have been a widely useful machine learning method for practical data analysis. Although they are extensively applied to time series data too, the features of dependence in the data are usually ignored, or too complex to be well-considered in resampling. In this paper, we suggest some new statistical ideas to build random forests for time series modelling, which is called Random Forests by Random Weights (RF-RW). The proposed method avoids the difficulty or complexity of using bootstrap resampling techniques for time series data by assigning random weights for the observations. It thus overcomes the shortcoming of generating bootstrapped resamples that lead to the breakdown of the time order and dependence structure of original time series data. Asymptotic properties and numerical results are developed for the proposed RF-RW. Application to COVID-19 analysis is demonstrated.

Session 35	**High-dimensional Regression and Applications** *Organizer/Chair:* *Gaorong Li, Beijing Normal University* *Jingxuan Luo, Beijing Normal University*	E22- 2011	11 Dec, 16:30 - 18:00

Probabilistic exponential family inverse regression and its applications

Daolin Pang, Shanghai Jiao Tong University

Rapid advances in high-throughput sequencing technologies have led to the fast accumulation of high-dimensional data, which is harnessed for understanding the implications of various factors on human disease and health. While dimension reduction plays an essential role in high-dimensional regression and classification, existing methods often require the predictors to be continuous, making them unsuitable for discrete data, such as presence-absence records of species in community ecology and sequencing reads in single-cell studies. To identify and estimate sufficient reductions in regressions with discrete predictors, we introduce probabilistic exponential family inverse regression (PrEFIR), assuming that, given the response and a set of latent factors, the predictors follow one-parameter exponential families. We show that the low-dimensional reductions result not only from the response variable but also from the latent factors. We further extend the latent factor modeling framework to the double exponential family by including an additional parameter to account for the dispersion. This versatile framework encompasses regressions with all categorical or a mixture of categorical and continuous predictors. We propose the method of maximum hierarchical likelihood for estimation, and develop a highly parallelizable algorithm for its computation. The effectiveness of PrEFIR is demonstrated through simulation studies and real data examples.

Data thinning for Poisson factor models and its applications

Zhijing Wang, Shanghai Jiao Tong University

The Poisson factor model is a powerful tool for dimension reduction and visualization of large-scale count datasets across diverse domains. Despite the availability of efficient algorithms for estimating factors and loadings, existing methods either require prior knowledge of the number of factors, or resort to ad hoc criteria for its determination. This paper proposes a novel data-driven criterion called Information Criterion via Data Thinning (ICDT), leveraging the thinning property of the Poisson distribution. Unlike traditional data splitting, data thinning partitions the count matrix into training and validation sets while preserving both the distribution and the underlying data structure. Interestingly, the validation error can be decomposed into the apparent error plus a covariance penalty. A simple estimator of the covariance penalty is obtained, leading to the development of ICDT. The selection consistency of ICDT is derived when both the sample size and the number of variables diverge to infinity. The proposed methodology is extended to dimension reduction in regression by incorporating the response inversely into the Poisson factor model. Extensive simulated examples and two real data applications are used to evaluate the performance of ICDT and compare it with existing criteria.

Calibrated Equilibrium Estimation and Double Selection for High-dimensional Partially Linear Measurement Error Models

Jingxuan Luo, Beijing Normal University

In practice, measurement error data is frequently encountered and needs to be handled appropriately. As a result of additional bias induced by measurement error, many existing estimation methods fail to achieve satisfactory performances. This paper studies high-dimensional partially linear measurement error models. It proposes a calibrated equilibrium (CARE) estimation method, calibrating the bias caused by measurement error and overcoming the technical difficulty of the objective function unbounded from below in high-dimensional cases due to non-convexity. To facilitate the applications of CARE estimation method, a bootstrap approach for approximating covariance matrix of measurement *errors is introduced. For the high dimensional or ultra-high dimensional partially linear measurement error models, a calibrated equilibrium multiple double selection (CARE-MUSE) algorithm, a novel multiple testing method, is suggested to control the false discovery rate (FDR) of significant covariates. We obtain the oracle inequalities for prediction risk and estimation error and the bound of the number of falsely discovered signs for the CARE estimator under some regularity conditions. The convergence rate of the estimator of the nonparametric function is also established. FDR and power guarantee for CARE-MUSE algorithm are investigated under a weaker minimum signal condition, which is insufficient for the CARE estimator to achieve sign consistency. Extensive simulation studies and an actual data application demonstrate the satisfactory finite sample performance of the proposed methods.*

A kernel independent test using projection-based measure in high-dimension

Yuexin Chen, Renmin University of China

Testing the independence between two high-dimensional random vectors is a fundamental and challenging problem in statistics. Most existing tests based on distance and kernel may fail to detect the non-linear dependence in the high-dimensional regime. To tackle this obstacle, this paper proposes a kernel independence test for assessing the independence between two random vectors based on a class of Gaussian projections relying on tuning parameters. The proposed test can be generally implemented for a wide class of distance-based kernels and completely characterizes dependence in the low-dimensional regime. Besides, the test captures pure non-linear dependence in the high-dimensional regime. Theoretically, we develop central limit theorem and associated rate of convergence for the proposed statistic under some mild regularity conditions and the null hypothesis. Moreover, we derive the asymptotic power of the proposed test enabling us to select suitable parameters for a special alternative, to achieve superior power in the high-dimensional regime. The choices of tuning parameters ensure that the proposed test has comparable power with the original kernel-based test in the moderately high-dimensional regime. Numerical experiments also demonstrate the satisfactory empirical performance of the proposed test in various scenarios.

Session 36	**Statistical Inference for Complex Data** *Organizer/Chair:* *Huazhen Lin, Southwestern University of Finance and Economics* *Jiakun Jiang, Beijing Normal University*	E22-2013	11 Dec, 16:30 - 18:00

Online Differentially Private Inference for Linear Regression Model

Xuerong Chen, Southwestern University of Finance and Economics

In the era of big data, data privacy has attracted increasing attention. Differential privacy(DP is a state-of-the-art framework for formal privacy guarantees. Many privacy-preserving inference methods have been developed for releasing information from a wide range of data analyses in the DP framework. However, DP statistical inference methods for streaming data, which represent a common type of big data, are still lacking. In this paper, we propose a computationally efficient privacy-preserving method for online update and inference of linear regression models that satisfies DP guarantees. We derive regression parameter estimates in the DP framework, along with the covariance estimates based on which privacy-preserving confidence intervals for the parameters are constructed. We provide theoretical support for the proposed DP method, and numerical results demonstrate the good performance of our approach.

Minimax Detection Boundary and Sharp Optimal Test for Gaussian Graphical

Bin Guo, Southwestern University of Finance and Economics

In this paper, we derive the minimax detection boundary for testing a sub-block of variables in a precision matrix under the Gaussian distribution. Comparing to the results on minimum rate of signals for testing precision matrices in literature, our result gives the exact minimum signal strength in a precision matrix that can be detected. We propose a thresholding test which is able to achieve the minimax detection boundary under certain cases by adaptively choosing the threshold level. The asymptotic distribution of the thresholding statistic for precision matrices is derived. Power analysis is conducted to show the proposed test is powerful against sparse and weak signals. Simulation studies show the proposed test has an accurate size around the nominal level, and it is more powerful than the existing tests for detecting sparse and weak signals in precision matrices. A real data analysis on brain imaging data is carried out to illustrate the utility of the proposed test in practice, which reveals functional connectivity between brain regions for Alzheimer's disease patients and normal healthy people.

Nonparametric Testing for Homogeneity with Alpha-Divergence in Reproducing Kernel Hilbert Spaces

Fode Zhang, Southwestern University of Finance and Economics

The statistical test method based on kernel mean embedding has attracted much attention due to its broad applicability. This paper proposes nonparametric homogeneity testing by embedding probability distributions on the reproducing kernel Hilbert space (RKHS). The alpha-divergence and its variational representation are employed to measure the discrepancy of the distributions. The Rademacher complexity bound of test statistics is discussed. The asymptotic distribution of the statistics under the null hypothesis is obtained. The power analysis and minimax optimality under different alternatives are investigated. We find that the performance of the test can be improved by tuning the parameter alpha. The proposed method can be applied to test the data with various structures. As applications, the method is used to test the homogeneity of degradation paths. The Monte Carlo simulation study and LED degradation data analysis are reported to illustrate the effectiveness of the test method suggested in this paper.

Sample efficient nonparametric regression via low-rank regularization

Jiakun Jiang, Beijing Normal University

Nonparametric regression suffers from curse of dimensionality, requiring a relatively large sample size for accurate estimation beyond the univariate case. In this paper, we consider a simple method of dimension reduction in nonparametric regression via series estimation, based on the concept of low-rankness which was previously studied in parametric multivariate reduced-rank regression and matrix regression. For d > 2, the low-rank assumption is realized via tensor regression. We establish a faster convergence rate of the estimator in the (approximate) low-rank case. Limitations of the model are also discussed. Through simulation studies and real data analysis, we compare the estimation accuracy of the proposed method with that of existing approaches. The results demonstrate that the proposed method yields estimates with lower RMSE compared to existing methods.

Session 37	**Accounting and Business Analytics** *Organizer/Chair: Jason Xiao, University of Macau;* *Morris Liu, University of Macau*	E22-2014	11 Dec, 16:30 - 18:00

Peer Effects in Corporate Investment Based on Product Networks Using BERT Model

Jianing Li, Northeastern University

The traditional method of industry classification is insufficient for the study of corporate relationships. Firms have broken through the competition in the same industry and launched cross-industry investment and diversification around new products. Products with similar characteristics can form the basis of product networks, which connect different firms. Firms modify their investment strategies in response to interactions with other firms within the product network. This paper proposes a novel approach to calculating product-text similarity for redefining and quantifying peer effects in corporate investment based on BERT model. This paper identifies peer firms and calculates product similarity by textually analysing the product composition of main business revenue disclosed in the notes of the annual reports of Chinese A-share listed companies from 2009 to 2022.

The empirical results show that there is a positive correlation between investment decisions among firms based on product networks. Although the difference between product similarity among product peers within the same industry and that of product peers cross-industry is slight, focal firms may focus more on competitors within the industry. When the variance of the focal firm's gross product margin is higher, the focal firm's investment is positively related to the product peers' investment; when the focal firm's net working capital is higher, the focal firm's investment is negatively related to the product peers' investment. Investment efficiency affected by product peers' investment is negatively and insignificantly related to the return on investment of the focal firm. However, after splitting the sample, investment efficiency affected by product peers within the industry is positively related to the focal firm's return on investment, and investment efficiency affected by product peers cross-industry is negatively and insignificantly related to the focal firm's return on investment. The results indicate that instead of chasing trends and investing, firms should focus on deepening their expertise in their product niche. This paper provides a new methodology and empirical support for understanding the complex interactions between firms.

Model construction and scenario application of big language model in ESG information analysis

Wenyi Li, Central China Normal University

The aim is to utilize emerging technologies such as big language models and big data to expand, accelerate, and enhance the acquisition and analysis of ESG information, in order to achieve automation, digitization, intelligence, democratization, and transparency in ESG information analysis. [Method/Process] Based on the big language model, a universal and personalized ESG information analysis model is constructed. The module functions and technical implementation process of ESG information analysis are elaborated, and the application of the ESG information analysis model in different entities such as external investors, internal enterprises, and government regulatory agencies is realized. The applicability, feasibility, and effectiveness of the big language model in ESG information analysis are verified through case studies. [Results/Conclusion] The ESG information analysis model based on the big language model can improve the efficiency and effectiveness of ESG information analysis, empower stakeholders' subjective initiative, achieve personalized and intelligent ESG information application scenarios, and make ESG information analysis as transparent and democratic as possible. [Originality/Value] Based on the theory of data value chain, this article systematically, comprehensively, and completely elaborates on the automation of the entire process of ESG data *collection, storage, processing, analysis, and application from the perspective of data life cycle. It provides more intelligent ESG information analysis tools for different entities in the capital market, which is beneficial for investor decision-making, enterprise self-improvement, and government regulatory agencies to supervise the healthy operation of the market.*

The Impact of Exercising Rights by Regulatory Minority Shareholder on Management Discussion and Analysis (MD&A) Tone: Evidence from China

Jason Xiao / Zhenxin Li, University of Macau

We explore the impact of bank competition, measured by the geographical density of banking branches, on the readability of firms' annual reports. Using the sample of 41,295 firm-year observations of listed non-financial firms in China between 2007 and 2022, we find that intensified bank competition can significantly prompt firms to present their annual reports in less readable formats. Moreover, the effect is more pronounced when firms are managed by managers with higher manipulation incentives and when firms are far from optimistic about their performance. However, the reduced impact of banking competition on the readability of firms' annual reports becomes less apparent when these firms face a decreased probability of securing bank loans and when their annual reports are subject to intensified scrutiny. Additionally, our findings remain consistent with a series of endogeneity and robustness tests, such as exogenous policy shocks, an instrumental variable, and alternative measures. Taken together, the results of our empirical analysis indicate that the readability of firms' annual reports decreases with the increased competition in the banking industry.

Bank Competition and annual report readability

Morris Liu / Yue Li, University of Macau

We explore the impact of bank competition, measured by the geographical density of banking branches, on the readability of firms' annual reports. Using the sample of 41,295 firm-year observations of listed non-financial firms in China between 2007 and 2022, we find that intensified bank competition can significantly prompt firms to present their annual reports in less readable formats. Moreover, the effect is more pronounced when firms are managed by managers with higher manipulation incentives and when firms are far from optimistic about their performance. However, the reduced impact of banking competition on the readability of firms' annual reports becomes less apparent when these firms face a decreased probability of securing bank loans and when their annual reports are subject to intensified scrutiny. Additionally, our findings remain consistent with a series of endogeneity and robustness tests, such as exogenous policy shocks, an instrumental variable, and alternative measures. Taken together, the results of our empirical analysis indicate that the readability of firms' annual reports decreases with the increased competition in the banking industry.

Session 39	**Structure Learning for High-dimensional Complex Data** *Organizer/Chair: Jingyuan Liu, Xiamen University*	E22- 2015	11 Dec, 16:30 - 18:00

Simultaneous Dimension Reduction and Variable Selection for Multinomial Logistic Regression

Canhong Wen, University of Science and Technology of China

Multinomial logistic regression is a useful model for predicting the probabilities of multiclass outcomes. Because of the complexity and high dimensionality of some data, it is challenging to fit a valid model with high accuracy and interpretability. We propose a novel sparse reduced-rank multinomial logistic regression model to jointly select variables and reduce the dimension via a nonconvex row constraint. We develop a block-wise iterative algorithm with a majorizing surrogate function to efficiently solve the optimization problem. From an algorithmic aspect, we show that the output estimator enjoys consistency in estimation and sparsity recovery even in a high-dimensional setting. The finite sample performance of the proposed method is investigated via simulation studies and two real image data sets. The results show that our proposal has competitive performance in both estimation accuracy and computation time.

Change-Points Detection and Support Recovery for Spatiotemporal Functional Data

Decai Liang, Nankai University

Large volumes of spatiotemporal data, including patterns of climatic variables, satellite images and FMRI data, usually exhibit inherent mean changes. Due to the complicated cross-covariance structure, the full covariance function is commonly described as a product of independent spatial covariance and temporal covariance, which is a mathematically convenient yet not always reflective assumption of the data. To remedy this, we propose a novel hypothesis test based on a more realistic assumption known as weak separability. We establish solid asymptotic theory to support this approach. Furthermore, we develop a comprehensive procedure for support recovery amidst the intricate correlations between space and time, effectively identifying true signals (locations with mean change) while controlling the false discovery rate. This represents the first work of support recovery within a spatiotemporal framework. Simulation studies and a Chinese precipitation data application validate the efficacy and enhanced power of our methodology on both change point detection and support recovery.

Dynamic Matrix Recovery

Ying Yang, Fudan University

Matrix recovery from sparse observations is an extensively studied topic emerging in various applications, such as recommendation system and signal processing, which includes the matrix completion and compressed sensing models as special cases. In this article, we propose a general framework for dynamic matrix recovery of low-rank matrices that evolve smoothly over time. We start from the setting that the observations are independent across time, then extend to the setting that both the design matrix and noise possess certain temporal correlation via modified concentration inequalities. By pooling neighboring observations, we obtain sharp estimation error bounds of both settings, showing the influence of the underlying smoothness, the dependence and effective samples. We propose a dynamic fast iterative shrinkage-thresholding algorithm that is computationally efficient, and characterize the interplay between algorithmic and statistical convergence. Simulated and real data examples are provided to support such findings. Supplementary materials for this article are available online.

Consistent Estimation of Structural Break Models with Endogenous Regressors

Chuang Wan, Jinan University

This article develops a novel approach to consistently identify the number of break points in structural break models with endogeneity. We propose a group LASSO procedure combined with a block segmentation scheme to quickly identify a set of possible break points. This step is practically easy to implement and can be performed efficiently. We derive the theoretically optimal tuning parameter for the LASSO estimation. Since practitioners are less willing to select the tuning parameter in practice, we suggest directly choosing the best subset from the candidate break point set based on a predetermined information criterion. However, the performance of this approach depends on the specified penalty factor and the optimal magnitude usually varies from the model and error distribution. To address this issue, we further develop a sophisticated cross-validation criterion incorporating an order-preserved sample-splitting strategy tailored for change point models. This method ensures selection consistency under some mild conditions. For testing purposes, we develop a score-type test statistic as an alternative to the Wald-type statistics. Intensive simulation studies reflect the promising aspects of our methodologies.

Session 40	**Novel Statistical Methods for Complex Data** *Organizer/Chair: Kai Kang, Sun Yat-sen University;* *Qingzhi Zhong, Jinan University*	E22- 2017	11 Dec, 16:30 - 18:00

An automatic MDDM-based test for martingale difference hypothesis

Guochang Wang, Jinan University

Checking whether the error term is a marginal difference sequence (MDS) in the multivariate time series model with a parametric conditional mean is a crucial problem.Tests based on the martingale difference divergence matrix (MDDM) are an effective statistical method for testing MDS in the residuals of multivariate time series models. However, MDDM-based tests require specifying the lag order. To solve this problem, we propose a data-driven MDDM-based test that automatically selects the lag order.This method has three main advantages:first, researchers do not need to specify the lag order while the test automatically selects it from the data;second, under the null hypothesis, the lag order is one, which significantly reduces computational costs;third, the proposed automatic tests have good performance in detecting model inadequacy caused by high-order dependence.In theory, we prove the asymptotical property of the proposed method. Furthermore, we demonstrate the effectiveness of this method through simulations and real data analysis.

The nonparametric GARCH model estimation using intraday high-frequency data

Xingfa Zhang, Guangzhou University

Most of nonparametric GARCH models typically employ daily frequency data to forecast the returns, correlations, and risk indicators of financial assets, without incorporating alternative frequency data. As a result, valuable financial market information may remain underutilized during the estimation process. To partially mitigate this issue, we introduce the intraday high-frequency data to enhance the estimation of the volatility function in a nonparametric GARCH model. To achieve this objective, we introduce a nonparametric proxy model for volatility. Under mild assumptions, we derive the asymptotic bias and variance of the estimator and further investigate the impact of various volatility proxies on estimation accuracy. Our findings from both simulations and empirical analysis indicate a considerable improvement in the estimation of the volatility function through the introduction of high-frequency data.

Bayesian Analysis of ARCH-M model with a dynamic latent variable

Zefang Song, Guangzhou University

A time-varying coefficient ARCH-in-mean (ARCH-M) model with a dynamic latent variable that follows an AR process is considered. The joint model extends the existing ARCH-M model by considering a dynamic structure of latent variable for examining a latent ef- fect on the time-varying risk–return relationship. A Bayesian approach coped with Markov Chain Monte Carlo algorithm is developed to perform the joint estimation of model pa- rameters and the latent variable. Simulation results show that the proposed inference pro- cedure performs satisfactorily. An application of the proposed method to a financial study of the Chinese stock market is presented.

High-dimensional covariate-augmented overdispersed poisson factor model

Qingzhi Zhong, Jinan University

The current Poisson factor models often assume that the factors are unknown, which overlooks the explanatory potential of certain observable covariates. This study focuses on high dimensional settings, where the number of the count response variables and/or covariates can diverge as the sample size increases. A covariate-augmented overdispersed Poisson factor model is proposed to jointly perform a high-dimensional Poisson factor analysis and estimate a large coefficient matrix for overdispersed count data. A group of identifiability conditions are provided to theoretically guarantee computational

identifiability. We incorporate the interdependence of both response variables and covariates by imposing a low-rank constraint on the large coefficient matrix. To address the computation challenges posed by nonlinearity, two high-dimensional latent matrices, and the low-rank constraint, we propose a novel variational estimation scheme that combines Laplace and Taylor approximations. We also develop a criterion based on a singular value ratio to determine the number of factors and the rank of the coefficient matrix. Comprehensive simulation studies demonstrate that the proposed method outperforms the state-of-the-art methods in estimation accuracy and computational efficiency. The practical merit of our method is demonstrated by an application to the CITE-seq dataset.

Session 52	**Financial Market, FinTech, and Information Disclosure** *Organizer/Chair: Wenjin KANG, University of Macau*	E22-2018	11 Dec, 16:30 - 18:10

ChatGPT, Stock Market Predictability and Links to the Macroeconomy

Jian Chen, Xiamen University

We find that good news extracted by ChatGPT from the front pages of Wall Street Journal can predict the stock market and is related to macroeconomic conditions. Consistent with existing theories, investors tend to underreact to positive news, especially during periods of economic downturns, high information uncertainty and high novelty of news. In contrast, the negative news is only associated with contemporaneous returns. Traditional methods of textual analysis, such as word lists and large language models like BERT, can barely find any predictability. In short, ChatGPT appears the best AI in discerning economic-related news that drive the stock market.

Does Media Sentiment on Target Innovation Predict Prosperous Technology Mergers and Acquisitions?

Yugang Chen, Sun Yat-sen University

Media plays a crucial intermediary role in disseminating innovations and arising the public awareness. But little is known in extant literature about how to measure media sentiment of innovation and its impact on corporate investments like acquiring technology via mergers and acquisitions (M&As). In this study, we self-construct a text-based measure for media sentiment of innovation (MedSI) reflecting a target's innovation productivity as reported in news released around the M&A announcements and find it predicts superior short-term post-announcement performance in Chinese technology M&As. We further dissect MedSI into its expected component (MedPred) and unexpected component (MedBias) and reveal that the positive impact on short-term post-announcement performance is primarily attributed to MedPred, especially when overall market investor sentiment or attention is strong. Moreover, MedPred demonstrates the ability to predict superior long-term post-announcement performance and innovation productivity. These findings remain robust in identification checks and various tests involving subsamples and alternative measures.

The Economics of Aisle Crossing: Depolarizing the US Congress

Bo Li, Peking University

We exploit exogenous variation in Senator and Representative seating locations to find novel evidence of the depolarizing potential of seating locations. Namely, we find that exogenously assigned seating assignments of Senators and Representatives have sizable impacts on the manner in which both their beliefs evolve and on their voting behavior - moving both toward more depolarizing stances when exposed to such. We exploit the seniority rank and assignment system of US Senate Chamber seating, along with the House Lottery system for the assignment of incoming freshmen legislators' offices to identify legislators' exogenous exposures to forced peers. We use these exogenously placed politicians over the last 30 years and over 1 million votes to show that their choices and voting behavior are profoundly impacted by their 'randomly assigned' neighboring legislators. For instance, a one standard deviation increase in the percentage of randomly assigned 'forced neighbors' in the Senate that vote yes (or no) on a bill increases the probability of the exogenously placed legislator of voting in that same direction – above and beyond party, ideology, state-level and other motives - by 7 percentage points (t=6.47). These effects are even larger on: close bills, bills that are less important to the given Senator's economic interests, bills from the most polarized bill topics, and as the Senator spends more time (gets more exposure) to the peer group. These results hold both within and across party-lines, with moderate members of each party exerting the largest impact toward depolarization.

The Impact of Market Competition on Donation-Based Crowdfunding: Mediation Effects of Information Disclosure

Fan Li, Shenzhen University

The rise of donation-based crowdfunding in recent years has resulted in intensifying competition among charitable organizations for donations. This study utilizes a unique dataset compiled from all fundraising campaigns on a leading crowdfunding platform and employs a panel vector autoregressive model to examine the dynamic relationships among competition, information disclosure, and fundraising. Our mediation analysis shows that, although market competition has a negative direct effect on fundraising amounts, it indirectly increases amounts through enhanced information disclosure by charities, leading to an insignificant total effect of market competition. The indirect effect is particularly notable at the campaign initiation stage, where increased transparency leads to higher donations. In contrast, reduced disclosure of ongoing project progress also results in increased fundraising. Our research contributes to a broader understanding of donation-based crowdfunding by illuminating the interplay between competition, strategic information disclosure, and donor behavior, offering insights into how charities can navigate competitive environments to optimize fundraising outcomes.

Cultural Value in Digital Arts

Fai Lim Loi, Macau University of Science and Technology

Do collectors value the culture in artwork? We formalize a stylized auction model that connects the owner's subjective valuation of artwork to its optimal price offering strategy, bids from potential buyers, and transaction prices of artwork. We test our hypothesis with a detailed on-chain auction history of CryptoPunks, a pioneering non-fungible token project featuring punk culture. Tokens with punk features are 3.6 ETH more expensive, equivalent to a 5.6% premium. Consistent with model predictions, punk token owners seek higher returns in the offer prices and receive more bids before they agree to sell; thus, punk tokens are less frequently traded.